Robustness, Reproducibility and Ecological Consistency in the Demarcation of Operational Taxonomic Units from Complex Sequencing Data

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Thomas Sebastian Benedikt Schmidt

aus Deutschland

Promotionskomitee

Prof. Dr. Christian von Mering (Vorsitz)

Prof. Dr. Kentaro Shimizu

Prof. Dr. Wolf-Dietrich Hardt

Zürich, 2014

"I wish to God," said Gideon with mild exasperation, "that you'd talk – just once – in prose like other people."

- Dorothy Dunnett, *The Game of Kings*

Table of Contents

1. Summary

The microbial world is notoriously elusive to direct observation. Microbes are '*small and many*', and studying them in a community context is a formidable challenge, both technically and conceptually. Technical challenges have traditionally resided in *resolution* and *throughput*, but are arguably being overcome by recent advances in sequencing technology. However, while detailed microbial surveys are becoming available for many environments, technological bias remains an issue, as organisms are being observed only indirectly, represented by sequences. Moreover, conceptually, a mere census is little more than a 'parts list' of an environment and not necessarily informative of the ecological roles of organisms, nor of their interactions. Even more pressing conceptual challenges reside in the lack of a unifying bacterial species concept, and in the identification of meaningful microbial diversity units from complex sequencing data. In practice, 'true' microbial lineages are often approximated by *Operational Taxonomic Units* (OTUs), defined as clusters of sequence similarity with respect to a taxonomic marker gene. Although OTUs are arguably '*proxies for proxies of proxies*' (lineages are approximated as clusters of marker gene sequences, which in turn represent organisms), they are an integral part of the contemporary microbial ecology toolbox.

The demarcation of 'meaningful' OTUs from complex sequencing datasets remains an open problem. Many approaches to sequence clustering have been suggested, but in spite of notable attempts towards increased standardization, no universally applied one-fit-all method has emerged. Rather, the choice of sequence clustering method introduces variability when analyzing microbial ecology data. In this thesis, one main aim has been to quantify this variability introduced by the choice of OTU demarcation method, and to assess the impact of method choice on downstream biological descriptions. In a multidimensional approach, OTUs were demarcated from a global, comprehensive dataset of small subunit (SSU) rRNA gene sequences according to different widely employed clustering methods, and under varying clustering parameters. The analyses revealed surprising trends in the similarity of partitions in terms of cluster composition, as well as in the robustness of methods to changing parameters. The presented results pertain to the *reproducibility* of biological findings in microbial ecology: they explore how robust OTU-based analyses are to the choice of experimental approach.

In a complementary analysis, it was investigated how well impartially clustered OTUs approximate 'true' microbial lineages. One frequently cited criterion for 'good' (i.e., theory-compliant) units of microbial diversity is *ecological consistency*. The general ecological consistency of OTUs was assessed based on curated contextual sequence annotations. It was found that OTUs are indeed generally, though not perfectly, ecologically consistent, at least at the studied ecological resolution. However, there were marked differences in ecological consistency between different widely used methods. As ecological similarity is generally correlated with SSU sequence similarity, and as ecological homogeneity is a criterion for 'true' microbial lineages, the observed differences in OTU ecological consistency were interpreted in terms of clustering *quality*.

The findings presented in this thesis may inform the design of microbial ecology studies, and recommendations on the choice of clustering method are provided. Moreover, the presented findings are potentially relevant beyond microbial ecology, in particular to the fields of microbial taxonomy and systematics. As the study of microbial communities advances, analyses such as presented in this thesis will be integral to providing robust, reproducible and consistent approaches to the computational analysis of complex sequencing data.

2. Zusammenfassung

Die Welt der Mikroorganismen entzieht sich der unmittelbaren Beobachtung. Mikroben sind '*klein und zahlreich*' – sie im Kontext von Lebensgemeinschaften zu untersuchen ist eine technische und konzeptionelle Herausforderung. Technische Herausforderungen stellten traditionell *Auflösung* und *Durchsatz* dar; jedoch scheinen diese durch jüngste Fortschritte der Sequenzierungstechnologie grösstenteils überwunden. Nichtsdestotrotz bedingt die indirekte Beobachtung von Mikroorganismen – welche durch spezifische Sequenzen lediglich repräsentiert werden – weiterhin technologieimmanente Bias. Des Weiteren entspricht selbst ein umfassender Zensus mikrobiellen Lebens in einem bestimmten Mikrobiom letztlich nur einer 'Liste von Bestandteilen' mit begrenzter Aussagekraft bezüglich ökologischer Zusammenhänge. Wesentlich dringlichere konzeptionelle Hürden stellen zudem der Mangel eines einheitlichen theoretischen Unterbaus für bakterielle Spezies-Konzepte, sowie die sinnvolle Unterteilung komplexer Sequenzdatensätze in elementare Diversitätseinheiten dar. In der Praxis werden mikrobielle Spezies häufig durch *Operational Taxonomic Units* (OTUs, etwa: 'operationelle Taxonomieeinheiten') angenähert, definiert als Gruppen ('Cluster') von Markergen-Sequenzen mit hoher Ähnlichkeit. Obwohl OTUs unbestreitbar '*stellvertretende Stellvertreter von Stellvertretern*' sind (Spezies werden als Cluster von Markergen-Sequenzen angenähert, die wiederum stellvertretend für Organismen stehen), sind sie unverzichtbar im Arsenal zeitgenössischer Methoden der mikrobiellen Ökologie.

Die Unterteilung komplexer Sequenzdatensätze in 'sinnvolle' OTUs ist dabei ein ungelöstes Problem. Viele Ansätze existieren, jedoch ist trotz grosser Anstrengungen hinsichtlich stärkerer Standardisierung keine einzelne Methode als universell anwendbar und sinnvoll etabliert. Stattdessen bedingen unterschiedliche Ansätze eine erhöhte Variabilität in der Analyse ökologischer Datensätze. Ein Hauptaugenmerk der vorliegenden Dissertation liegt auf der Quantifizierung dieser Flexibilität, die durch die Wahl unterschiedlicher OTU-Definitionen erzeugt wird, sowie auf deren Einfluss auf nachfolgende biologische Analysen. In einem mehrdimensionalen Ansatz wurde ein globaler Datensatz von *small subunit* (SSU) ribosomalen RNA-Gensequenzen in OTUs unterteilt, gemäss einiger weit verbreiteter OTU-Definitionen und unter wechselnden Parametern. Mehrere überraschende Beobachtungen bezüglich Ähnlichkeit von OTU-Sets zwischen Methoden, sowie bezüglich der Anfälligkeit von Clustering gegenüber Parametervariation werden beschrieben. Die diskutierten Ergebnisse betreffen insbesondere die *Reproduzierbarkeit* biologischer Erkenntnisse in der mikrobiellen Ökologie: sie erkunden, wie robust OTU-basierte Analysen gegenüber unterschiedlichen experimentellen Zugängen sind.

In einem komplementären Ansatz wurde zudem untersucht, wie gut OTUs 'tatsächliche' mikrobielle Spezies anzunähern vermögen. Ein häufig erwähntes Kriterium für 'gute' (im Sinne von, 'theorie-konforme') Diversitätseinheiten ist *ökologische Konsistenz*. Diese wurde mithilfe von kuratierten Sequenz-Metadaten für OTUs bestimmt. Es konnte gezeigt werden, dass OTUs in der gewählten ökologischen Auflösung generell, jedoch nicht vollkommen, ökologisch konsistent sind. Es wurden jedoch deutliche Unterschiede zwischen verschiedenen Methoden beobachtet. Da ökologische und SSU-Sequenz-Ähnlichkeit korrelieren, und da ökologische Homogenität ein Merkmal 'tatsächlicher' mikrobieller Spezies ist, lassen sich die beobachteten Unterschiede in der ökologischen Konsistenz als Qualitätsunterschiede interpretieren.

Die in dieser Dissertation präsentierten Ergebnisse tragen zur informierten Planung und Durchführung von Experimenten in der mikrobiellen Ökologie bei, betreffen darüber hinaus aber auch die mikrobielle Taxonomie und Systematik. Mit zunehmendem technologischen Fortschritt in der Untersuchung mikrobieller Lebensgemeinschaften können Ansätze wie der hier präsentierte zur Entwicklung robuster, reproduzierbarer und konsistenter Ansätze zur Analyse komplexer Sequenzdatensätze beitragen.

3. Introduction

3.1 The small subunit rRNA gene in studies of microbial ecology

Since its introduction as a phylogenetic marker in the late 1970ies [1,2], the small subunit ribosomal RNA (SSU rRNA) gene has become a cornerstone of modern microbiology. Functionally, the SSU rRNA molecule is an essential component of the ribosome: it scaffolds ribosomal proteins, contains the *Shine-Dalgarno* sequence [3] and stabilizes codon-anticodon binding during translation [4]. These functions are conveyed by an intricate folding of the ~1,500nt long molecule into a characteristic secondary and tertiary structure, where regions of specific base-pairing (hairpin loops) alternate with less structurally defined segments (Figure 3.1). This close coupling of SSU structure to ribosome function puts potent and site-specific constraints on the gene's evolution: nine '*hypervariable*' regions alternate with more conserved segments [5]. Nevertheless, given that ribosome function has been conserved for at least 3.8 billion years, it has been argued that the SSU rRNA gene evolves overall *uniformly*: it is subject to a largely unchanged selection regime, one of the reasons why it has been considered an 'Ultimate Molecular Chronometer' [6]. In this view, SSU divergence is seen as an indicator of phylogenetic relationships, because it reduces the 'noise' of any evolutionary processes that may go beyond mere sequence divergence over time, such as e.g. changing selective pressures leading to adaptive evolution of specific genome parts. Although various doubts have been raised regarding the utility of the SSU gene for reconstructing phylogenies [7,8], it remains the most widely used marker gene in studies of prokaryotic taxonomy and phylogeny on various scales [9-11].

At the same time, SSU rRNA-based approaches have revolutionized the research field of microbial ecology. By definition, microbial ecology is the study of microbes in context of their biotic and abiotic environment. For this, microbial ecologists had traditionally relied on direct observation, either by microscopy or from enrichment cultures (as pioneered by Beijerink, [12]), as well as on more indirect, biochemical evidence (e.g., in Winogradsky's classical column gradient experiment [13]). However, with the rise of molecular biology, marker gene-based approaches – in particular those relying on the SSU rRNA gene – have drastically changed experimental approaches to microbial ecology problems. In this regard, marker gene sequences sampled from an environment of interest serve as proxies for the underlying community of organisms, such that community-level ecological parameters (e.g., measures of diversity) can readily be deduced.

3.1.1 Four main assets of SSU-based approaches to microbial ecology

But what makes SSU sequencing so particularly attractive to microbial ecologists? There are arguably four main answers to this question, although indeed not all of them are specific to the SSU gene. First, SSU-based approaches are *cultivation-independent*. In 1985, Lane et al first described a protocol to sequence SSU rRNA without prior isolation and cloning of the gene, thus side-stepping the need to cultivate microorganisms in order to study them [14]. While the isolation and cultivation of (uncharacterized) microbes in the laboratory remains cumbersome and often impossible, it also introduces a biased perspective on microbial communities: 'cultivability' does not necessarily correlate with abundance in a community [15,16]. For example, while *Escherichia coli* is readily cultivable, it constitutes less than 0.1% of the healthy human gut microbiome, together with other facultative anaerobes [17]. The cultivation-independent study of (microbial) diversity by direct sequencing of environmental samples circumvents such problems; it is generally referred to as *metagenomics* [18-20].

Second, the SSU rRNA gene is comparatively well-studied: it is a cornerstone of microbial taxonomy and systematics, as detailed above. In 1994, Stackebrandt & Goebel [10] first introduced a SSU sequence similarity signal into bacterial species definitions: they observed that organisms from the same bacterial 'species', as

defined based on DNA-DNA hybridization kinetics [21], shared ≥97% SSU sequence similarity. Indeed, since 2002, official descriptions of new bacterial or archaeal taxa "should include an almost complete 16S rDNA sequence" [22]. Generally, although the definition of 'true' microbial lineages based on SSU data remains an open problem [9], a dependable SSU-based taxonomic framework is available, allowing to map uncultured diversity onto a taxonomic namespace (e.g., [23]). Moreover, there is an equally rich *theoretical* framework: as many approaches to phylogeny reconstruction have traditionally relied on SSU data [6,24,25], comprehensive SSU-based reference phylogenies are available, as are methods for reconstructing phylogenies directly from (environmental) SSU sequence data.

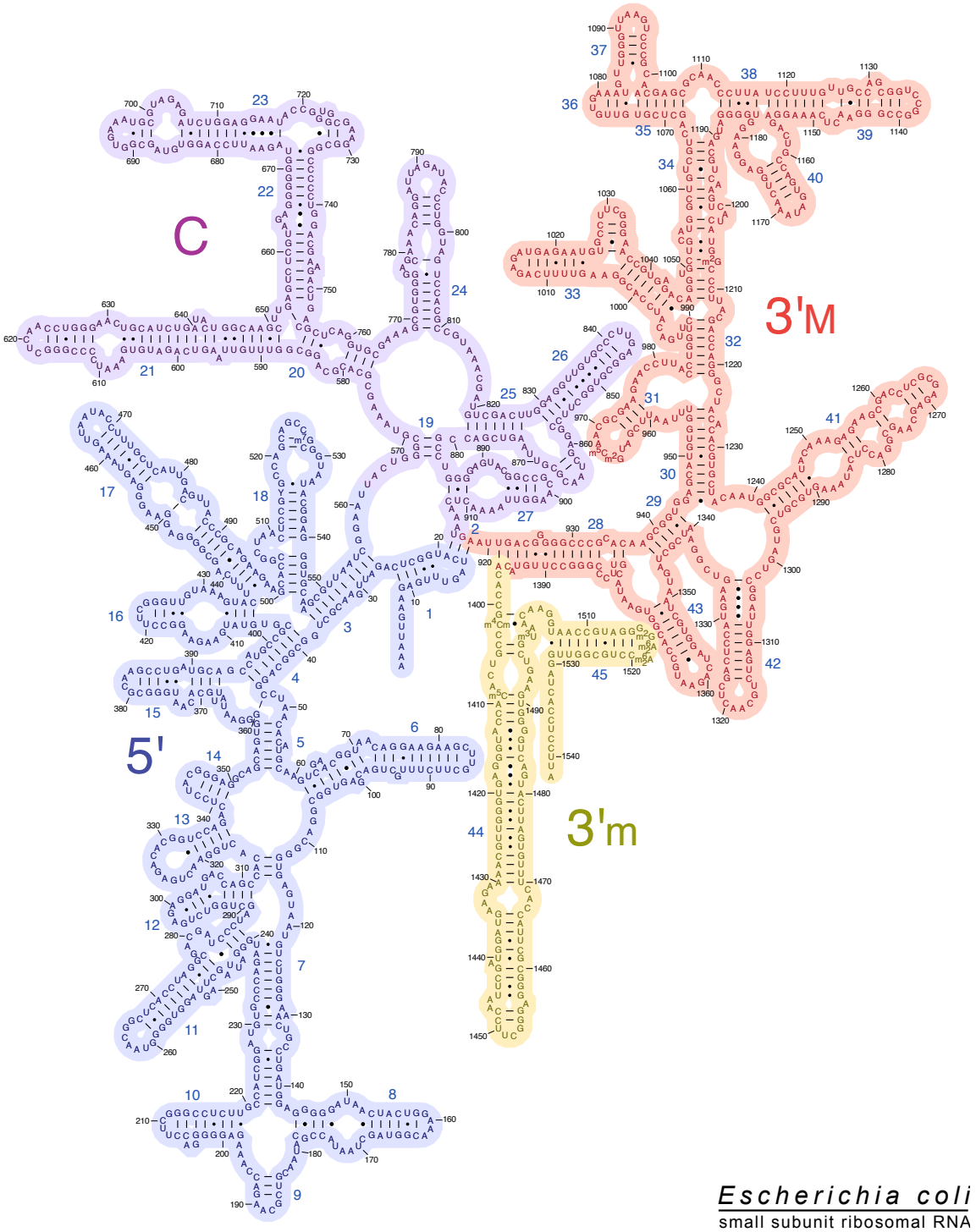


Figure 3.1. Secondary structure of the E. coli 16S rRNA. Image courtesy by Harry Noller, University of California, Santa Cruz.

Third, SSU sequence data is abundant in dedicated repositories. In 1992, Woese and coworkers first introduced the *Ribosomal Database Project* (RDP) which provides a collection of available rRNA gene sequences, as well as reference alignments and phylogenies [26,27]. Other databases, such as *Greengenes* [28] and *SILVA* [23,29] have since provided complementary sequence collections. As of March 2014, *RDP*[*] contains 2,929,433 SSU rRNA sequences, *Greengenes*[†] holds 1,262,986 SSU sequences and *SILVA*[‡] lists 4,058,383 sequences. All three repositories cover broad phylogenetic ranges and provide reference alignments, taxonomies and phylogenies, as well as additional tools for SSU-based studies. Thus, they provide rich context and valuable references when analyzing SSU datasets.

Fourth, advances in sequencing technology have enabled the study of individual environments at very high resolutions. Sequencing platforms such as the *Roche 454* [30] or *Illumina* [31] systems may provide millions of reads for an individual sample, although such high throughput comes at the cost of shorter read lengths and higher error rates when compared to 'traditional' Sanger sequencing [31,32]. For SSU-based approaches, these limitations have in part been overcome by (i) the design of dedicated primers which specifically target one or more hypervariable SSU subregions, and (ii) by the introduction of stringent sequencing noise filters (see section 3.2.2). In consequence, the microbial diversity in several environments has been studied to great depths: the *human microbiome project* (HMP) alone provided almost 50 million SSU reads from subregions V13 and V35 [34], and the *earth microbiome project*[§] [35] lists almost 1.9 billion SSU reads as of March 2014.

Thus, high-resolution SSU-based studies of microbial communities are technically attainable (due to advances in cultivation-independent sequencing protocols) and potentially highly informative (due to a rich taxonomic and phylogenetic background and large reference databases). While the microbial world has traditionally been elusive to direct observation, targeted metagenomics has provided insights into microbial community composition and dynamics of diverse environments, at various levels.

3.1.2 Major challenges to SSU-based approaches in microbial ecology, and how they may be alleviated by sequence clustering

At the same time, there remain several challenges to SSU-based approaches in microbial ecology. First, although cultivation-independent techniques reduce sampling bias, they do not altogether eliminate it. For example, PCR primers used for targeted amplification of (partial) SSU sequences from isolated DNA are often specific to certain taxa of interest, but even general-purpose primers may show (slight) binding preferences. As such minor differences are amplified during PCR cycles, they may drastically distort observed abundance patterns [36,37].

Second, PCR amplification may also introduce another type of artifacts, known as *chimeric sequences*. Formed by 'template-switching' of the DNA polymerase during amplification, PCR chimeras are artificial sequences composed of segments from two or more parental molecules that were present in the original sample. Several studies have quantified the penetrance of PCR chimeras in targeted sequencing datasets, but absolute estimates of chimera frequencies vary (e.g., [38-40]). Generally, however, studies concurred that chimeric sequences are

[*] http://rdp.cme.msu.edu

[†] http://greengenes.secondgenome.com

[‡] http://www.arb-silva.de

[§] http://www.microbio.me/emp/

abundant in many datasets and in consequence, dedicated chimera filtering methods have been developed (e.g., [40,41], see section 3.2.3).

Third, increased sequencing throughput comes at the cost of higher levels of *sequencing noise*. Although per-base errors for different sequencing technologies are in the low (sub-)percent range [42-45], the introduced artifacts are potentially detrimental for SSU-based analyses that rely on individual nucleotide differences to achieve fine-scale taxonomic resolution. The impact of sequencing noise may to some extent be alleviated by increased coverage and through dedicated 'de-noising' algorithms (e.g., [46,47]), but erroneous base calling remains a problem in targeted sequencing approaches [48]. Moreover, the choice of sequencing technology introduces other sources of platform-specific bias [49].

Fourth, high-throughput sequencing also imposes limitations on read length: current generation sequencing platforms may provide maximum read lengths of ~800-1,000nt, but in most use cases, attainable lengths are ≤200nt (Illumina) to ≤600nt (Roche 454) in practice. Thus, as the SSU gene spans ~1,500nt in total, current platforms do not allow sequencing of the full-length molecule. Rather, sequencing effort is usually targeted to selected hypervariable SSU subregions; for example, the *human microbiome project* targeted subregions V1-V3 and V3-V5 in complementary sequencing protocols [34]. One obvious drawback of shorter reads is information sparsity: compared to the full-length molecule, they provide a lower information content (number of nucleotides) and thus potentially lower resolution. Moreover, different hypervariable regions evolve at different rates [5], and findings based on different SSU subregions are often not trivially portable [50].

Finally, there remain important challenges regarding the downstream computational analysis of complex SSU sequencing datasets. From a technical point of view, the mere quantity of data is overwhelming: individual sequencing runs may provide millions of reads. In 2009, Grice et al [51] published a dataset of ~120,000 full-length 16S sequences sampled from different human skin sites; at the time, this was one of the largest environmental Sanger sequencing datasets available. Only three years later, in June 2012, the *human microbiome project* dataset was published, comprising almost 50 million Roche 454-sequenced 16S reads, sampled in a large, long-term collaborative effort [34]. As of March 2014, the *earth microbiome project* lists individual studies with tens of millions of Illumina-sequenced reads. In other words, sequencing technology is outpacing Moore's law (see Figure 3.2; [52]), and computer power has struggled to keep up.

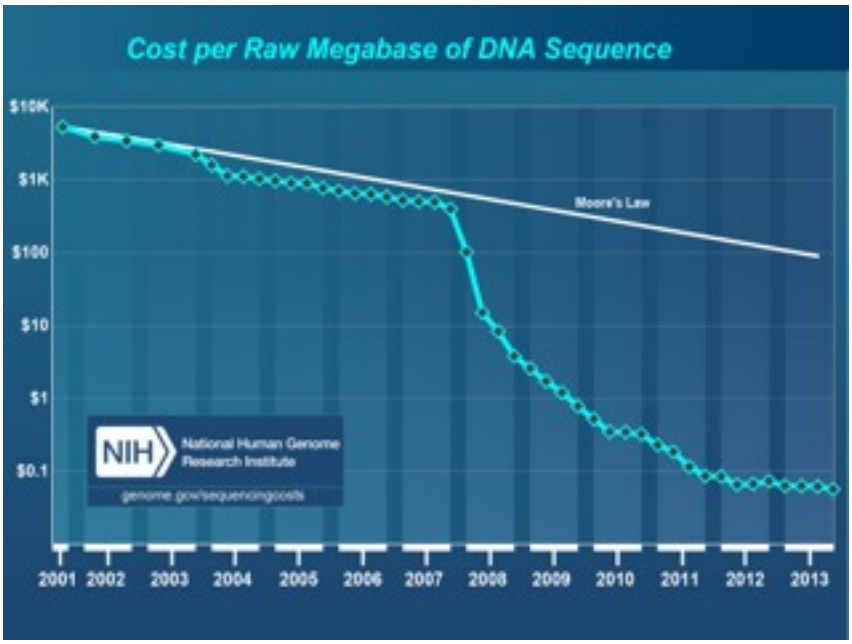


Figure 3.2: DNA sequencing outpaces Moore's law. The plot shows the evolution of per-megabase sequencing cost, but trends for the evolution of throughput per experiment are proportional. Image from [52], http://www.genome.gov/sequencingcosts/, accessed on March 28th, 2014.

Sequence clustering is one approach to conquer such datasets of increasing size computationally. Clustering by sequence similarity may minimize dataset redundancy [49], control for sequencing noise [53] and reduce dataset complexity to computationally accessible scales for downstream analyses. Moreover, as marker gene similarity indicates phylogenetic relatedness, clusters of SSU sequence similarity indeed carry biological meaning – they correspond to groups of closely related organisms. In practice, such similarity clusters may thus approximate microbial taxa at different levels; they are often referred to as *Operational Taxonomic Units* (OTUs).

The use of OTUs may therefore alleviate certain problems associated with SSU-based studies of microbial communities – both regarding technical difficulties (by reducing dataset complexity) and conceptual challenges (by approximating clusters of phylogenetic relatedness). However, OTU demarcation in its own right is also associated with a set of open questions. Many approaches to OTU binning have been employed, implementing different assumptions on the fundamental organization of microbial diversity, but a biologically 'optimal' OTU definition has remained elusive. Moreover, many parameter choices are available at different steps during OTU demarcation pipelines, but their impact on clustering variability and reproducibility is often not well understood. Likewise, although marker gene similarity carries a signal of phylogenetic relatedness, the degree to which OTUs represent 'true' microbial taxa is a matter of controversial debate. The work presented in this thesis revolves around these and other conceptual challenges associated with OTU demarcation.

3.2 Demarcating Operational Taxonomic Units to conquer complex sequencing datasets

The concept of OTUs was originally introduced by Sneath & Sokal in the 1960ies, in the context of *numerical taxonomy* [54]. By 'Operational Taxonomic Unit', Sneath and Sokal generally referred to a *group of entities being studied*, corresponding to individuals, species, genera, etc depending on the chosen resolution. The definition of 'OTUs' in context of microbial ecology is more specific: they are proxies for microbial taxa, demarcated based on marker gene similarity [55]. Taxonomic resolution depends on the choice of sequence similarity threshold when defining OTUs: the most widely adopted threshold is arguably ≥97% SSU sequence similarity, which corresponds to a traditional 'species-level' cutoff [10,56]. Indeed, several authors have used the term 'OTU' as synonymous surrogate for microbial 'species' [55].

3.2.1 Phylotyping, reference-based OTU binning and *de novo* OTU demarcation

There are two general approaches to partitioning complex SSU sequencing datasets; the supposed main distinction is the degree to which they infer additional biological information to the system under study via an explicit taxonomic mapping of sampled SSU sequences [57,58]. In *taxonomy-dependent* (or '*phylotyping*') approaches, SSU sequences are identified with close representatives from taxonomically annotated databases, e.g. using dedicated tools such as the *RDP Classifier* [59]. Such studies profit from a potentially rich biological context for the SSU data at hand: taxonomic names carry biological meaning both *semantically* (many taxa designations are indicative of phenotypic or ecological properties) and *historically* (they will mean *something* to other researchers in the field). However, labeling strongly depends on the coverage and annotation quality of the reference taxonomy used, and many microbial taxa remain poorly studied [16] or underrepresented in public repositories due to biased sampling efforts [49].

Alternatively, in *taxonomy-independent* approaches, sequences are binned into OTUs, either by *de novo* clustering, or by *reference-based* OTU picking. In the latter case, sequences are mapped to a set of reference OTUs, usually clustered from global, comprehensive databases as provided e.g. by *RDP* [27], *Greengenes* [28] or

SILVA [23]. In consequence, *reference-based* OTU binning potentially enhances portability of findings across studies, if common sets of reference OTUs are used. However, similarly to phylotyping, reference-based approaches heavily depend on the quality of the reference that is used. In contrast, *de novo* OTU demarcation does not invest prior knowledge: sequences are clustered impartially, based on dataset-internal signals alone. However, this potentially increases sensitivity to sequencing errors and other artifacts, such as chimeric sequences. The work presented in this thesis focuses mainly on *de novo* OTU clustering which is arguably more widely adopted than phylotyping and reference-based OTU binning.

In general, established protocols for the *de novo* demarcation of OTUs from complex SSU sequencing datasets follow five conceptual steps: raw sequences are (i) filtered to remove *sequencing noise* and (ii) *chimeric reads*, and subsequently (iii) aligned; from the alignment, (iv) sequence distances are computed and used to (v) cluster reads into OTUs. At each step, a choice of different methods is available, but there have been great efforts to standardize analysis workflows: dedicated '*one-stop*' pipelines such as *mothur* [60] or *QIIME* [61] provide wrapper scripts and informed defaults for existing software tools. Nevertheless, as the choice between different tools is largely left to each researcher's arbitration, there is significant methodological variability between published SSU-based analyses. In the next sections, common options at each step are highlighted, with a clear focus on the different proposed approaches to sequence clustering.

3.2.2 Filtering sequencing noise

Different sequencing platforms introduce characteristic types of sequencing noise, due to distinctive technological and chemical constraints, and noise profiles differ in PCR-based targeted sequencing versus whole genome sequencing approaches [48]. Traditional Sanger sequencing is considered largely 'noise-free', at estimated per-base error rates of $\sim 10^{-5}$ [62]. The Roche 454 platform is prone to erroneous base-calling for (long) homopolymers, resulting in artificial insertions or deletions [42,63]. In 2009, Quince et al introduced *PyroNoise*, an algorithm that corrects for such errors by pre-clustering the raw light intensity flowgrams prior to base calling [46]. One year later, Reeder & Knight introduced *Denoiser*, which relies on the same principle, but significantly reduces computational effort by pre-filtering flowgrams based on abundance distributions [64]. In 2011, Quince et al introduced *AmpliconNoise*, an update of *PyroNoise* which combines fast alignment-free flowgram clustering with a sequence pre-clustering step [47]. As of March 2014, *AmpliconNoise* and *Denoiser* are the most widely adopted denoising algorithms for Roche 454 pyrosequencing datasets; they are also the default choices in *QIIME* [61].

Illumina sequencing produces largely unspecific sequencing errors, but provides per-base read quality scores ('*Phred* scores', [62]) on which most noise filtering approaches rely. In 2012, Bokulich et al proposed an approach which is now widely adopted in Illumina-based SSU sequencing: they filtered reads by per-base quality scores, truncating reads at positions of marked drops in read quality [48].

In addition, platform-independent noise filtering approaches based on sequence pre-clustering have been proposed, e.g. by Huse et al, 2010 [53]. An in-depth discussion of sequencing noise filtering exceeds the scope of this thesis; comparative benchmarks of available methods have been provided e.g. by Schloss et al, 2011 [65] and Bonder et al, 2012 [66].

3.2.3 Removal of chimeric sequences

PCR chimeras are artificial sequences generated by template-switching of the DNA polymerase during PCR amplification (see section 3.1.2). Several methods to remove chimeric reads from sequencing datasets have been proposed, including *Bellerophon* [67], *Pintail* [39], *Ccode* [68], *Perseus* [47], *Uchime* [41] and *ChimeraSlayer* [40]. More recently, Edgar (2013) introduced *uparse* which combines sequence clustering with on-the-fly filtering for chimeric sequences based on abundance patterns [69]. As of March 2014, *ChimeraSlayer* is the default chimera detection method in *QIIME*, while *mothur* relies on *Uchime*. Comparative benchmarks of selected chimera detection methods can be found in [65] and [66].

3.2.4 Sequence alignment and distance calculation

Two main approaches to sequence alignment have been advocated for OTU demarcation approaches: *pairwise sequence alignment* (PSA), which optimizes alignments for pairs of sequences, and *multiple sequence alignment* (MSA), which optimizes alignments for entire sequence sets. The most widely used PSA algorithm was proposed by Needleman & Wunsch in 1970 [70]; it is based on dynamic programming, computationally efficient and trivially parallelizable [71]. Software tools such as *ESPRIT* and *ESPRIT-Tree* rely on the Needleman-Wunsch algorithm [71,72], which the authors of these tools have motivated in several comparative studies [58,73].

In contrast, MSA is computationally more demanding, in particular for large datasets – optimal MSA is a NP-complete problem [74-76] – but it incorporates dataset-wide information on sequence homology [58]. Although several computationally efficient MSA implementations such as *MUSCLE* [77,78] or *MAFFT* [79] are available, SSU sequencing dataset scopes are usually prohibitive of full MSA. Rather, many approaches rely on (curated) reference alignments, provided e.g. by *RDP, Greengenes* and *SILVA*, to which query sequences are aligned individually [60]. More recently, model-based MSA approaches have received increasing attention. In 2009, Nawrocki et al introduced *Infernal*, which aligns SSU sequences to curated covariance profiles of the SSU rRNA molecule that include secondary structure information [80,81]. Thus, *Infernal* provides structure-informed pseudo-MSA in a *many-to-one* approach (alignment to a common reference model) which is trivially parallelizable. As of March 2014, it is the default method implemented in *mothur* [60] and by *RDP* [27].

Once an alignment is achieved, many options are available to calculate pairwise sequence distances. Generally, sequence distance is the number of differences between two sequences divided by the entire alignment length (number of alignment columns). However, 'mismatches' may be defined differently, in particular in the treatment of alignment gaps, and specific substitution models may incorporate assumptions on sequence evolution into distance calculation. Three frequently used distance calculators are the '*one gap*' (gaps of any length counted as one mismatch), '*each gap*' (each gap position counted as mismatch) and '*no gap*' (alignment gaps ignored) calculators [50]. An in-depth discussion of alignment methods and distance calculation exceeds the scope of this thesis; comparative studies are available e.g. in refs [50,58,73,82-84].

3.2.5 OTU demarcation

A plethora of methods for the *de novo* demarcation of OTUs has been proposed; Table 3.1 provides a non-exhaustive overview. Conceptually, available approaches fall into five general categories: (i) *hierarchical clustering algorithms* (HCAs), (ii) hierarchical clustering with heuristic pre-partitioning (*hybrid* methods), (iii) *greedy incremental* or *seed-based heuristic* HCA approximations, (iv) *soft-threshold* or *threshold-free* recursive partition optimization methods and (v) algorithms relying on additional signals, other than mere sequence similarity. Clearly, these categories are neither rigid nor exclusive, and several available tools may fall into overlapping classes.

3.2.5.1 Hierarchical clustering algorithms

The most 'classical' approach to sequence clustering, adopted also in other disciplines of biology and beyond, are *hierarchical clustering algorithms*. Based on a full matrix of pairwise sequence distances, HCAs incrementally merge individual nodes (sequences) into clusters, starting with the most similar ones. While clustering to 0% similarity (the so-called *singleton partition*) and 100% similarity (filtering for unique sequences) is trivial, different *linkage methods* are available at non-trivial clustering thresholds. A linkage method is the set of rules according to which data is incrementally partitioned, and different linkage methods impose conceptually distinct clustering regimes.

The three most widely used linkage methods in *de novo* OTU demarcation are *complete, single* and *average* linkage (see Figure 3.3, top panel). During *complete linkage* (*cl*, or *furthest neighbor*) clustering, two clusters are merged if *all* pairwise similarities between cluster members are above the current similarity threshold. Thus, *cl* imposes an *exclusive* clustering regime: in the resulting partition, all pairwise similarities *within* clusters are guaranteed to be above the chosen threshold, but pairs of sequences sharing above-threshold similarity do not necessarily cluster together, depending on the succession of merging events. In other words, *cl* prevents *false positive* merges (sequence pairs of below-threshold similarity within the same cluster), but not *false negatives* (above-threshold pairs in different clusters).

In contrast, *single linkage* (*sl*, or *nearest neighbor*) is *inclusive*: two clusters are merged if *any* two members share above-threshold similarity. Thus, *sl* guarantees that every sequence pair of above-threshold similarity is clustered together (*false negatives* are prevented), but within clusters, pairwise similarities may be below-threshold (*false positives* may occur if two dissimilar sequences are connected by a 'chain' of sequence pairs sharing above-threshold similarity). Moreover, *sl* clustering is virtually *deterministic*: the resulting partition at a given clustering threshold is fully independent of the order in which sequences are processed.

Finally, *average linkage* (*al*, or *average neighbor*, or *Unweighted Pair Group Method with Arithmetic Mean, UPGMA*) provides a middle ground between the two: clusters are merged if the *average* similarity of all members is above-threshold. Thus, *al* neither guarantees that all sequence pairs of above-threshold similarity are merged (as in *sl*), nor that all cluster-internal similarities are above-threshold (as in *cl*). However, due to the regime of average similarity, it has been argued that *al* is more robust to both *false positives* and *false negatives* than *cl* and *sl* [57].

In 2005, Schloss et al introduced the *DOTUR* suite of software tools which provided the first HCA implementation for OTU demarcation purposes [56]. The *mothur*[**] suite, published in 2009 [60], provided an update to *DOTUR* and continues to be actively developed. Both have been widely used by microbial ecologists:

[**] http://www.mothur.org

as of March 2014, Thompson Reuters' *Web of Science* service[††] lists 1,422 and 1,614 studies citing *DOTUR* and *mothur*, respectively. However, both *DOTUR* and *mothur* do not scale well with dataset size: they require the full sequence distance matrix – which scales quadratically with the number of sequences – to be loaded in memory (*DOTUR*) or to be read on-the-fly from the hard disk (*mothur*). More recently, João F Matias Rodrigues in our group developed *hpc-clust*, a fully parallelizable HCA implementation which overcomes these limitations [85]. *Hpc-clust* distributes the sequence distance calculation to an arbitrary number of compute nodes in a computer cluster or multicore computer, where distances are locally sorted and sent to a 'master' node only at the time of clustering. Depending on dataset complexity, *hpc-clust* enables clustering of millions of (unique) sequences on a medium-sized computer cluster within a few hours wall time.

3.2.5.2 Hybrid heuristic / hierarchical algorithms

Another approach to scale HCAs to very large sequence sets is *heuristic pre-filtering* of sequence distances. Probably the first algorithm adopting such an approach was *ESPRIT*, introduced by Sun, Cai et al in 2009 [71]. Based on the observation that for many real-world datasets, only 1-5% of pairwise sequence similarities are in a range that is relevant to most microbial ecology applications (i.e., 90-100% sequence similarity), *ESPRIT* imple-

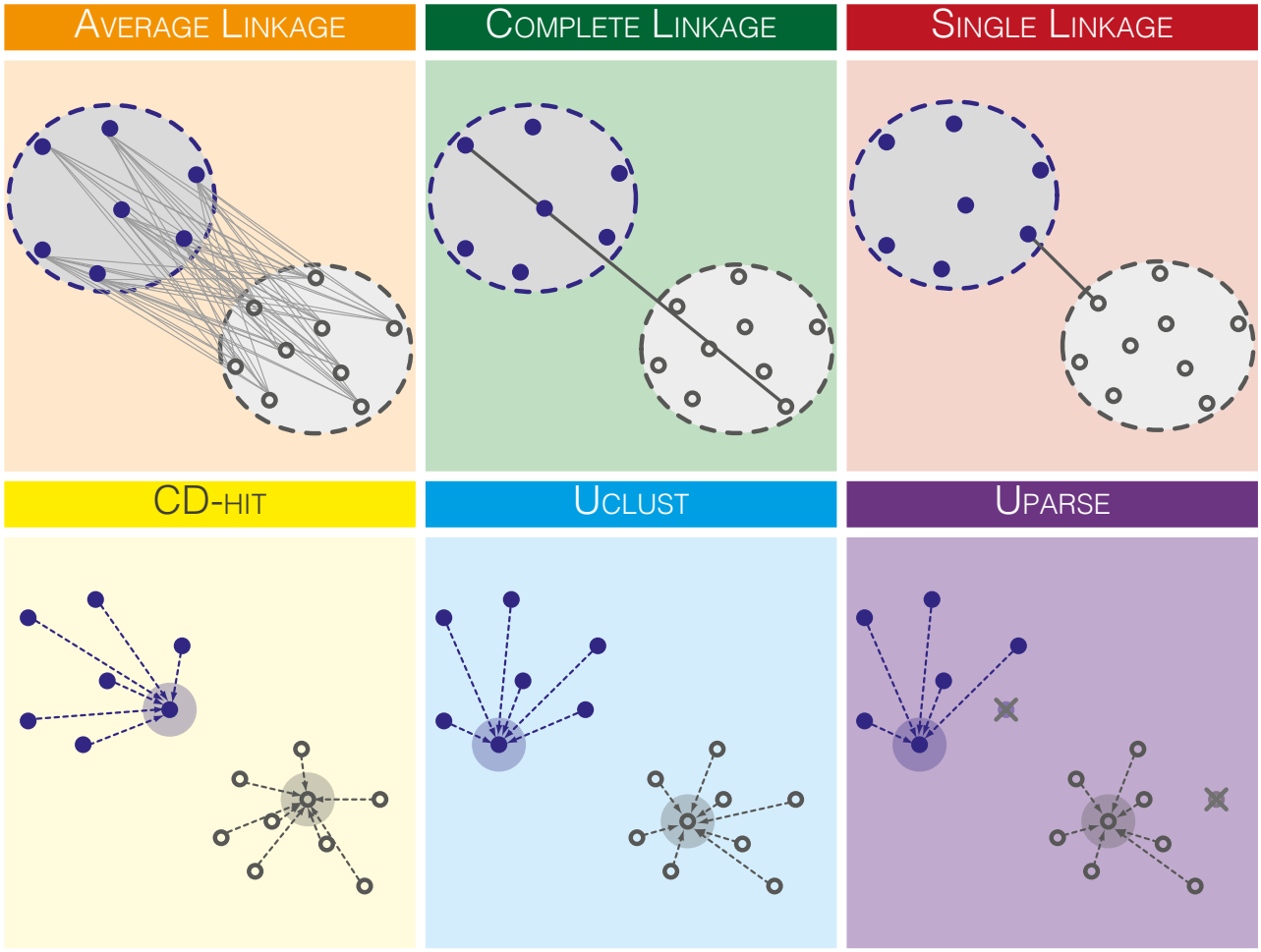


Figure 3.3: Overview of the six clustering methods studied in this thesis. Hierarchical clustering methods (top) rely on full sequence distance matrices, while heuristic methods (bottom) provide computationally efficient shortcuts, by implementing greedy seed-based clustering strategies.

[††] http://apps.webofknowledge.com

ments a fast *k-mer counting*-based exploration of the sequence distance space, filtering for 'relevant' pairwise distances to be calculated from full pairwise alignments. Subsequently, *cl* hierarchical clustering is performed based on this 'relevant' sequence subspace. In 2011, Cai & Sun introduced *ESPRIT-Tree*, which provided several additional refinements [72]. In a 'divide-and-conquer'-like approach, *ESPRIT-Tree* first heuristically pre-partitions the full sequence space into a '*pseudometric based partition tree*' and subsequently refines clustering in an *al*-like regime, including several more computationally efficient shortcuts. Both *ESPRIT* and *ESPRIT-Tree* provide parallel versions to run on computer clusters, and by design, they may provide significant speed improvements over full HCAs. However, in contrast to *DOTUR*, *mothur* and *hpc-clust* (which calculate sequence distances from MSAs), *ESPRIT* and *ESPRIT-Tree* implement on-the-fly PSA which entails a significant performance penalty. As of March 2014, *ESPRIT* and *ESPRIT-Tree* have received moderate attention from the microbial ecology community: the Thomson Reuters *Web of Science* service lists 92 and 24 studies citing these methods.

In 2010, Huse et al proposed heuristic *single linkage pre-clustering* (SLP) to 98% sequence similarity to pre-partition the full sequence space prior to hierarchical *al* clustering, primarily to reduce the effects of sequencing noise [53]. More recently, in 2012, Wei et al proposed *mBMK*, a '*modified bisecting k-means*' clustering algorithm [86]. In a two-step process, the sequence space is first pre-partitioned in an alignment free approach based on 'k-tuple' distances, followed by semi-supervised hierarchical k-means clustering (a top-down approach where the number of clusters is pre-selected).

3.5.2.3 (Greedy incremental) heuristic algorithms

While the aforementioned algorithms combine heuristic pre-partitioning steps with hierarchical clustering, many methods have been proposed that provide fully heuristic approximations of HCAs. One of the first heuristic sequence clustering methods was *cd-hit* (see Figure 3.3, lower left panel) which was originally developed to reduce redundancy in protein sequence databases [49,87-90]. Based on the observation that sequence similarity may be approximated by counting shared 'short words' (*k-mers*), *cd-hit* implements alignment-free *k-mer* counting based on *indexing tables*, which is computationally highly efficient. Sequence clustering is implemented in a *greedy incremental* algorithm, as first proposed by Hobohm et al in 1992 [91]: sequences are first pre-sorted (e.g., by length or abundance, depending on parameters), and the first sequence becomes the first *cluster seed*. Subsequent sequences are clustered to the seed if they share higher similarity than required by the chosen threshold, otherwise they form a novel cluster seed. This 'all-against-few' comparison strategy is computationally efficient and scales well with dataset size. However, clustering is potentially sensitive to the order in which sequences are processed, and the algorithm does not guarantee that close-matching sequence pairs cluster together. In 2012, Li et al introduced *cd-hit-otu*, a dedicated pipeline to demarcate OTUs from targeted high-throughput sequencing data [49]. Taken together, all *cd-hit* papers have been cited 1,610 times according to *Web of Science* (as of March 2014), but as *cd-hit* is a general purpose tool, many citing studies are not related to microbial ecology.

In 2010, Edgar introduced *uclust* (see Figure 3.3., central lower panel, [92]), which is a general purpose tool very similar to *cd-hit*. Conceptually, the main distinction between the two is that *uclust* implements actual PSA (in a BLAST-like approach called *usearch*) and by default clusters sequences to the *first* matching seed, rather than the *closest* match. Moreover, in more recent versions, *uclust* implements a different sequence similarity calculation function ('identities per alignment columns') than *cd-hit* ('identities per shorter sequence length'). *Uclust* is the default clustering method in *QIIME* and as of March 2014, it has been cited by 523 studies according to *Web of Science* (not all citing studies being related to microbial ecology).

Class	Ref	Name	Description
hierarchical clustering algorithms	[56,60], [85]	*Average Linkage (al)*	clusters merged if *average* similarity between all sequences exceeds the clustering threshold. *synonyms*: average neighbor, unweighted pair group method with arithmetic mean (UPGMA)
		Complete Linkage (cl)	clusters merged if *all* members share above-threshold similarity ('*exclusive*' clustering). *synonym*: furthest neighbor
		Single Linkage (sl)	clusters merged if *any* two members share above-threshold similarity ('*inclusive*' clustering). *synonym*: nearest neighbor
hierarchical / heuristic methods	[71]	*ESPRIT*	like *cl*, but with pairwise alignment and (slightly heuristic) k-mer-based pre-filtering of distances.
	[72]	*ESPRIT-Tree*	like *al*, but with pairwise alignment and heuristic pseudometric-based pre-partitioning ('divide and conquer')
	[53]	*Single Linkage Preclustering (SLP)*	pre-clustering to 98% similarity in a seed-based *sl* approach, then hierarchical clustering (*al*).
	[86]	*modified Bisecting K-Means algorithm (BMK)*	alignment-free distance calculation, followed by modified k-means clustering (i.e., split partitions until *k* clusters are left).
heuristic (greedy-incremental / seed-based) methods	[49]	*cd-hit*	sequences are compared to cluster representatives (seeds); sequence distances calculated from shared *k-mers*, alignment-free (using indexing tables).
	[92]	*uclust*	sequences are compared to cluster representatives (seeds); BLAST-like distance calculation (PSA).
	[69]	*uparse*	like *uclust*, but with on-the-fly filtering for chimeric sequences.
	[93]	*Taxonomy-Based Clustering (TBC)*	pre-clustering to 100% identity (unique sequences), then successive merging to largest clusters by adaptive BLASTing.
	[94]	*Dynamic Seed-Based Clustering (DySC)*	sequences are compared to representatives; cluster seeds are optimized on-the-fly (dynamic seeding).
	[95]	*MSClust*	adaptive on-the-fly selection of multiple cluster seeds.
	[96]	*Locally Sensitive Hashing (LSH-Div)*	iterative seeding; heuristic pairwise sequence distance calculation
'soft' and threshold-free methods	[97]	*CROP*	unsupervised Bayesian clustering, based on a Gaussian mixture model
	[98]	*Bayesian Estimation of Bacterial Communities (BEBaC)*	alignment-free pre-clustering, Bayesian 'fine' clustering ('divide and conquer')
	[99]	*CLUSTOM*	alignment-free pre-clustering with subsequent 'overlap minimization' by pairwise sequence alignment
	[100]	*M-Pick*	graph-based approach, OTUs assigned based on modularity; sequence distances calculated from pairwise alignments
clustering based on additional signals	[101]	*PhylOTU*	mapping of reads to a reference alignment, then clustering by phylogenetic distance (tree-based)
	[82]	*VI-Cut*	hierarchical clustering, semi-supervised to match (taxonomic) data labels, based on *Variation of Information* (*VI*)
	[103, 104]	*Ecotype Simulation (ES, QuickES)*	highly parametric simulation of 'ecotypes' from sequence similarity and a phylogenetic tree, assuming the 'Stable Ecotype Model' of bacterial evolution
	[102]	*Evolutionary Placement on Phylogenetic Trees (EPA-PTP)*	open-reference approach, using evolutionary placement on a reference tree and a 'Poisson tree processes' model.
	[106]	*Distribution-Based Clustering (DBC)*	abundance differences of sequences across samples ('distributions') are used as an ecological signal to inform clustering

Table 3.1. Selected methods for de novo OTU demarcation.

More recently, in 2013, Edgar introduced *uparse* (Figure 3.3, lower right panel) which is a refined greedy algorithm that combines sequence clustering with on-the-fly filtering for chimeric sequences [69].

Several other heuristic sequence clustering algorithms have been proposed, including '*Taxonomy-Based Clustering*' ([93], adaptive BLASTing against an abundance-sorted database of unique sequences) and *DySC* [94], *MSClust* [95] and *LSH-Div* [96] (greedy algorithms that rely on adaptive on-the-fly seed optimization, see Table 3.1). However, these have not (yet) been widely adopted by microbial ecologists.

3.5.2.4 'Soft'-threshold and threshold-free methods

While greedy heuristic algorithms sacrifice accuracy to increase computational efficiency, several methods have been developed that aim to optimize partitions at the cost of higher computational demands. Based on the observation that depending on data complexity, dataset-wide 'hard' similarity cutoffs do not necessarily reflect the 'natural' data structure, so-called *soft-threshold* or *threshold-free* methods aim to provide partitions that are 'optimal' under different statistical data models. The most widely used soft-threshold method is arguably *CROP*, introduced by Hao et al in 2011 [97]. *CROP* implements iteratively refined unsupervised Bayesian clustering using Gaussian mixture models, rather than probabilistic sequences, as cluster representatives. *BEBaC*, introduced by Cheng et al in 2012, is another Bayesian clustering method, implementing alignment-free 'crude clustering' followed by Bayesian 'fine clustering' in a divide-and-conquer approach [98]. *CLUSTOM* implements an alignment-free exploration of the sequence space, followed by pre-clustering and subsequent iterative *overlap minimization* between adjacent clusters [99]. Finally, *M-Pick* implements *modularity-based* clustering in a graph-based approach: sequences are connected based on pairwise similarity, and subsequently clustered in iterative refinement to identify partitions that best represent data-inherent 'modularity' [100].

3.2.5.5 Clustering based on additional (non-sequence) signals

Several methods have been proposed that rely on signals other than 'mere' sequence similarity to amend OTU demarcation. One of the earliest such methods is *PhylOTU*, introduced by Sharpton et al in 2011 [101]. *PhylOTU* relies on reference alignment-derived probabilistic sequences to which query reads are aligned (i.e., a reference MSA is iteratively extended). Subsequently, a phylogenetic tree is inferred from the full alignment of reference and query sequences, and query sequences are clustered based on *phylogenetic distance*, rather than sequence similarity. Thus, *PhylOTU* is one of the few methods that may cluster (non-overlapping) SSU reads from non-targeted metagenomics studies. In a similar approach, *EPA-PTP* implements an *evolutionary placement algorithm* ('EPA') to map reads to a (curated) reference phylogenetic tree and subsequently uses a *Poisson tree processes* ('PTP') model to infer 'species' boundaries in an open-reference approach [102]. *Ecotype Simulation* (ES) is another phylogeny-informed method to identify 'true' microbial lineages from complex sequencing datasets [103,104]. Assuming the *Stable Ecotype Model* of bacterial speciation [105], *ES* and its heuristic approximation, *QuickES*, provide (highly) parametric evolutionary simulations, using a sequence alignment and a phylogenetic tree as input.

VI-cut, proposed by White et al in 2010 [82], implements an entirely different approach in which clustering is 'guided' by taxonomic classification of sequences. *VI-cut* implements semi-supervised hierarchical clustering, such that the *Variation of Information* (*VI*) of assigned taxonomic labels is optimized. Finally, *Distribution-Based Clustering* ('DBC'), introduced by Preheim et al in 2013 [106], relies on an *ecological* signal to inform sequence clustering. *DBC* implements a greedy heuristic algorithm (similar to *cd-hit* or *uclust*) in which sequences are compared both in terms of nucleotide sequence similarity and in similarity of abundance profiles across multiple available samples (based on the observation that similar abundance profiles may imply ecological similarity).

3.4 Assessing the 'goodness' of OTUs

In view of the myriad of available approaches to OTU demarcation, several comparative studies have sought to identify suitable methods with respect to different concepts of 'optimal' data partitions. Table 3.2 provides a non-exhaustive overview of selected works. Based on the chosen benchmark parameters, approaches fall into four general categories: (i) optimization of *total OTU counts* when clustering (known) datasets; (ii) *external* benchmarking, using taxonomy as 'ground truth'; (iii) *internal* benchmarking, based on binary classification tests; and (iv) optimization for other signals. These categories are overlapping, and many studies have relied on complementary benchmarking strategies.

3.4.1 Benchmarking based on total OTU counts.

In the (OTU-based) ecological description of microbial communities, the determination of community *richness* is a common first step. Local richness is the number of 'species' (in practice approximated by OTUs) present in a sampled community of interest. In consequence, many benchmarking studies have compared OTU demarcation methods based on total cluster counts – based on the realization that many OTU-based studies tended to overestimate 'true' community richness, e.g. due to PCR and sequencing errors [42,46,107]. The underlying rationale is that clustering can be deemed 'good' if it *reduces diversity overestimation* (e.g., [53]).

In 2009, Sun, Cai et al [71] observed that their proposed method *ESPRIT* (see above) provided systematically fewer OTUs than 'traditional' MSA-based hierarchical clustering as implemented in *DOTUR*. They relied on a dataset of ~340,000 V6 reads pyrosequenced from a mock community of 43 strains of known 16S sequence, published by Huse et al in 2007 [42], as well as on environmental samples from hydrothermal vents [108] and an air sample. Based on the same re-sequenced mock community V6 dataset, as well as four lower-complexity re-sequenced datasets, Huse et al in 2010 observed marked differences in total OTU counts between different methods when clustering to a nominal threshold of 97% sequence similarity [53]. In particular, they observed that their proposed strategy of 98% heuristic *single-linkage preclustering* (*SLP*, see above) with subsequent hierarchical *al* clustering based on PSA provided the best approximation of the 'expected' number of 43 OTUs. In 2011, Barriuso et al [83] analyzed the same V6 mock dataset, as well as a simulated dataset, a full-length environmental dataset and two Roche 454 model datasets introduced by Quince et al [46]. They found that *ESPRIT* and *cl* clustering provided the most accurate OTU count estimates. Finally, Huse et al's V6 mock dataset, as well as several simulated datasets, were also used in a 2013 large-scale comparison of many existing clustering approaches by Chen et al [109]. They found that *SLP* and *CROP* were most robust in predicted OTU counts, but that several methods *underestimated* 'true' diversity at 97% clustering (in contrast to the *overestimation* observed in previous studies).

In 2010, Schloss reported "effects of alignment quality, distance calculation method [...] and region" on SSU-based OTU demarcation [50]. From a dataset of 13,501 unique, curated and aligned full-length 16S sequences shared between the *RDP, Greengenes* and *SILVA* databases, Schloss extracted several commonly targeted hypervariable regions and compared resulting OTU counts under *cl* clustering to full-length partitions. He observed that the different variable regions provided poor approximations of full-length clustering behavior. Moreover, he quantified effects of alignment strategy (PSA vs MSA) and sequence distance calculation. In a similar approach, Kim et al (2011) extracted hypervariable subregions from the full-length 16S alignments provided by *RDP*, corroborating the finding that subregion-based OTU counts generally deviate from full length-based counts [110].

Ref	Dataset	Optimization criterion	Main findings
[71]	V6 sequences (Huse et al's mock community, ocean water & air samples); benchmarks on subsets of 10,000 sequences	OTU counts	PSA preferable to MSA; *ESPRIT* preferable to *DOTUR*
[82]	1,677 sequences (*RDP*), trimmed to V2-V4	Variation of Information (*VI*) against taxonomic ground truth; diversity estimates	*cl* preferable to *al* & *sl*; *VI-cut* preferable to all
[53]	several datasets of V6 sequences, re-sequenced mock communities	OTU counts	*sl*-preclustering followed by *al* best reduced overestimation of OTU counts
[50]	13,501 full-length sequences (*RDP*, *SILVA* & *Greengenes*); multiple subregions extracted	OTU counts; diversity estimates	choice of subregion and reference alignment influence OTU clustering
[83]	various datasets (simulated & re-sequenced)	OTU counts; 'output variability'	*ESPRIT, RDP* or *mothur* preferable, depending on parameters
[110]	full-length sequences (*RDP*); multiple subregions extracted	OTU counts; richness estimates	choice of subregion influences OTU clustering; subregions approximate full-length clustering differentially well
[72]	sets of 30,000 V2 sequences (Turnbaugh et al)	Normalized Mutual Information (*NMI*) against taxonomic ground truth	*ESPRIT-Tree* preferable to *ESPRIT*, *uclust* and *cd-hit*
[58]	sets of 30,000 V2 sequences (Turnbaugh et al); several datasets on other subregions	OTU counts; *NMI* against taxonomic ground truth; F-score (*precision* and *recall*)	*ESPRIT-Tree* preferable to *ESPRIT*, *mothur*, *uclust* and *cd-hit*
[57]	14,596 full-length sequences (*RDP*, *SILVA* & *Greengenes*); V13 and V35 extracted	Matthew's Correlation Coefficient	*al* preferable to *cl, sl, uclust, cd-hit,ESPRIT* and *BlastClust;* pre-clustering by taxonomic label (*RDP-Classifier*) may enhance accuracy
[66]	V57 mock & oral dataset	OTU counts; *NMI* & 'purity' against taxonomic ground truth	closed-reference OTU picking preferable; *uclust* and *ESPRIT-Tree* preferable out of *de novo* methods
[49]	1.1 million V6 sequences (Turnbaugh et al); 5M Illumina reads	OTU counts; computational efficiency (speed)	*cd-hit-otu* is fast and accurate
[104]	116,391 full-length sequences (human skin); 1,025 *hsp60* & 132 *psaA* sequences	Phylogenetic consistency (monophyly); ecological homogeneity (microhabitats)	*QuickES*-simulated 'ecotypes' preferable to any OTU-clustering method
[109]	V6 sequences (Huse et al's mock community); simulated datasets	OTU counts; *precision, recall*; NID-score	*CROP* and *sl*-preclustering generally preferable

Table 3.2. Selected studies that benchmarked OTU demarcation.

3.4.2 External benchmarking against taxonomic ground truth

Based on the realization that total OTU counts is a summary statistic and does not inherently reflect partition *structure*, several studies relied on *taxonomy* as an external 'ground truth' to benchmark OTU demarcation. The underlying rationale is that OTU taxonomy is (i) representative of recognized microbial lineages, (ii) indicative of phylogenetic, morphological of metabolic coherence and (iii) relevant to the practitioner.

In 2010, White et al [82] investigated a dataset of 1,677 full-length, taxonomically typed 16S sequences downloaded from *RDP* and assessed clustering 'quality' as *Variation of Information* (*VI*; see section 5.5.2) with

respect to taxonomic labels. They found that hierarchical *cl* clustering, but in particular their proposed semi-supervised method *VI-cut* (see above) provided partitions which best approximated the distribution of taxonomic labels.

In 2011, Cai & Sun used a dataset of ~1.1 million V2 sequences from the human gut (first provided by Turnbaugh et al in 2009 [111]) to compare their proposed algorithm *ESPRIT-Tree* (see above) to existing methods [72]. They inferred taxonomic labels for the full sequence set by BLASTing against the *RDP* reference database and subsequently processed subsets of 30,000 randomly selected sequences. For these, compliance with taxonomic ground truth was measured as *Normalized Mutual Information* (*NMI*, see section 5.5.2), which is mathematically closely related to the *VI* measure used by White et al. Based on *NMI* scores, Cai & Sun concluded that *ESPRIT-Tree* (an *al* approximation) outperformed their previous method *ESPRIT* (a *cl* approximation), as well as heuristic *cd-hit* and *uclust* clustering. In a follow-up study, Sun, Cai et al corroborated these findings in the same experimental setup, but adding discrimination against *mothur-al* and *mothur-cl*, as well as some general observations on hypervariable SSU subregions [58].

In 2012, Bonder et al compared several OTU demarcation protocols based on a V5-V7 targeted sequencing dataset of oral samples comprising ~500,000 sequences [66]. Based on taxonomic labeling by BLASTing against the *SILVA* database, they assessed taxonomic 'purity' of clusters, as well as *NMI* scores and complemented their findings by observations on total OTU counts. While they "could not identify a single best clustering algorithm", they observed that sequence pre-processing generally had a stronger impact on clustering outcome than the choice of clustering method. Finally, in their 2013 study (see previous section), Chen et al [109] relied on the *NMI*-related *NID* score to quantify compliance with taxonomic ground truth.

The abovementioned studies arguably share two fundamental drawbacks. First, the use of a taxonomic 'ground truth' in OTU demarcation is conceptually problematic, for various reasons. In general, the relevance of taxonomic categories such as *species, genus* or *family* remains highly debated for bacteria and archaea [9,112,113]. Among other things, this has led to the development of multiple curated reference taxonomies which conflict on many levels [28,114]. Moreover, many cases have been reported in which taxonomic labeling conflicts with SSU similarity clusters or clusters of ecological coherence (e.g., [115-119]). Furthermore, taxonomic coverage of microbial diversity in reference databases is very uneven, due to highly biased sampling effort [16,49]; in consequence, taxonomic 'ground truth' may only refer to 'known' diversity, but is less indicative with respect to 'uncharacterized' (or 'un-named') diversity.

Second, it has been shown by Vinh et al (2009) that *NMI* and *VI* scores produce shifting baseline values, depending on cluster counts [120]. As OTU demarcation protocols vary markedly in total cluster counts (see above), these measures are therefore arguably unsuitable to benchmark against (taxonomic) 'ground truth'.

3.4.3 Internal benchmarking based on binary classification tests

Based on the realization that the indicative value of *external* benchmarks against taxonomic 'ground truth' is questionable, several studies have compared methods in an *internal* benchmarking approach. As sequence clustering can be interpreted as a *binary classification* problem, OTU partitions can be described using methods from machine learning theory. For a pair of sequences, clustering can be considered *true positive* (*TP*) if both sequences share above-threshold similarity and pertain to the same OTU. *True negative* (*TN*) pairs share below-threshold similarity and are not clustered together. *False positive* (*FP*) pairs are clustered to the same OTU, although they do not share the required level of similarity. *False negative* (*FN*) share above-threshold similarity,

but are not clustered together; see also section 3.2.5.1. Based on *TP, TN, FP* and *FN* counts, clustering *precision* (fraction of relevant positive classifications), *recall* (or *sensitivity*, fraction of detected relevant positive classifications) and *specificity* (rate of correctly identified *true negative* classifications) may serve as 'internal' measures of clustering quality.

In 2011, Schloss & Westcott [57] calculated *Matthew's Correlation Coefficient* – which is a weighted summary statistic of *TP, TN, FP* and *FN* counts – to compare different clustering methods, based on a dataset of 14,956 full-length, aligned 16S sequences downloaded from *RDP*. They found that *al* clustering provided the highest internal clustering consistency. In 2012, Li et al [49] found that their proposed pipeline *cd-hit-otu* provided accurate OTU count estimates, as well as high specificity and sensitivity when clustering the Turnbaugh dataset of ~1.1 million sequences sampled from the human gut [111] or the mock community datasets provided by Quince et al [46,47]. Finally, Chen et al (2013) assessed precision and recall of clustering methods, in addition to OTU count-based and taxonomy-based benchmarking [109].

3.4.4 Optimization for other signals

Remarkably few studies have assessed OTU demarcation with regard to signals that to not rely on OTU counts, taxonomy or classification theory. Most notably, Koeppel & Wu (2013) investigated *phylogenetic* and *ecological* signals [104]. They relied on three specific datasets: (i) ~115,000 Sanger-sequenced 16S reads of the *human skin microbiome* dataset [51]; (ii) 1,025 *hsp60* sequences of the genus *Vibrio* [121]; and (iii) 132 *psaA* sequences of the genus *Synechococcus* [122]. They observed "extensive and pronounced paraphyly and polyphyly among OTUs" with respect to a maximum-likelihood phylogenetic tree of the large 16S dataset. Moreover, they observed "extensive ecological heterogeneity among OTUs", clustered to 97% sequence similarity using *uclust*, with respect to very fine-scale ecological descriptions for the *hsp60* and *psaA* datasets. Koeppel & Wu advocated that their proposed method *QuickES* did not suffer from the same drawbacks: while they did not process the large 16S dataset using *QuickES*, their method provided a far higher number of clusters for the small datasets than 97% *uclust* OTU clustering. This much more fine-grained partition aligned reasonably well with very fine-scale ecological sample labels (based on slight variations in associated marine particle sizes for *Vibrio* and 2-3°C temperature differences between hot spring *Synechococcus* habitats). However, habitat associations were predicted using the tool *AdaptML* [121]. Both *QuickES* and *AdaptML* rely on a (common) phylogenetic tree as input, so that their reasonable agreement is unsurprising. Moreover, both *QuickES* and *AdaptML* are highly parametric methods, and given the very small size of the two test datasets, there is a risk of (congruent) overfitting of the data, which Koeppel & Wu did not address.

A very different approach to assessing the 'goodness' of OTU clustering was implemented by Edgar, 2013 [69], who compared different OTU demarcation protocols based on the number of unidentified chimeric sequences (as well as based on total OTU counts). Based on his definition of 'true chimeric' sequences, he found that his proposed method *uparse* (whose distinctive feature is a novel approach to on-the-fly chimera filtering) outperformed the default pipelines of *mothur* and *QIIME*, which implement *al* and *uclust*, respectively.

To our knowledge, beyond these two studies, there have been no approaches to assessing OTU clustering 'quality' *independently* of taxonomy, total OTU counts or internal test statistics.

4. Objectives and Content of this Thesis

As detailed in the previous sections, the demarcation of OTUs from complex sequencing data is important in microbial ecology, but a commonly accepted, unifying OTU-binning strategy remains elusive. Several studies have compared sequence clustering methods, benchmarking for different concepts of 'optimality'. However, by design, such uni-dimensional benchmarks have missed important trends *between* methods: when clustering the same set of sequences, how similar are the partitions provided by different methods? As OTUs are often the backbone of further biological analysis, a quantitative understanding of the variability introduced by OTU clustering is arguably relevant to microbial ecologists. Moreover, when assessing OTU *quality*, one important aspect has so far remained underexplored: the *biological consistency* of the clustering.

In light of these considerations, the work presented in this thesis pertains to five main questions: (i) how similarly do different widely used clustering methods partition complex SSU datasets into OTUs?; (ii) how robust are methods to (slightly) changing parameters?; (iii) how do differences between methods influence biological interpretation?; (iv) do impartially clustered OTUs approximate 'true' (i.e., ecologically coherent) microbial lineages?; and (v) how can differences in ecological consistency between clustering methods be interpreted in terms of 'clustering quality'?

To approach these questions, we demarcated OTUs from a global, comprehensive dataset of available full-length SSU sequences, and from subsets thereof. For six different clustering methods, we generated OTU sets across a wide range of thresholds, and for varying clustering contexts and SSU sequence subregions. We conducted a series of computational experiments that comprehensively established pairwise similarities between clustering methods across changing clustering thresholds, and observed differential levels of robustness to other changing clustering parameters. Moreover, we quantified systematic biases in ecological data descriptions (diversity estimation) between methods. Our findings are summed up in a research manuscript, reprinted in this thesis as manuscript 7.1, which has been submitted for publication.

Moreover, we explored *ecological consistency* to indicate how well OTUs approximate 'true' microbial lineages. We observed that OTUs are generally, though not perfectly, ecologically consistent, but that different clustering methods provide differential levels of consistency. These findings are summed up in manuscript 7.2, which has been accepted for publication in *PLOS Computational Biology*.

The methods used for the reported experiments are detailed in section 5; in particular, this section also provides a critical methods discussion that complements the reprinted manuscripts. Section 6 highlights major results and puts them in context; it complements the discussions provided in the reprinted manuscripts, but does not replace them.

5. Methods

5.1 Dataset preparation

The majority of experiments discussed in this thesis were conducted on three types of SSU sequence datasets which were generated from publicly available data: (i) a *global* dataset containing roughly 1 million SSU sequences sampled from a wide range of environments; (ii) two *'local'* subsets of this larger set, notably the well-studied *Human Skin Microbiome* (HSM) dataset [51] and an artificially generated dataset of *broad ecological range* (BER); (iii) three simulated *short read* datasets, extracting SSU subregions V23, V35 and V6 from full-length sequences in the global set.

5.1.1 A comprehensive survey 16S-based survey of microbial diversity

When studying the demarcation of OTUs, there are various possible choices to define a test dataset. Many contemporary studies in microbial ecology rely on high-throughput sequencing technologies, such as the Roche 454 and Illumina platforms. These methods provide very deep sequencing at low costs, but are prone to sequencing noise and erroneous reads [46,47]. Moreover, different sequencing methods introduce different biases to data generation [45], and technological advance in terms of read length, sequencing depth and quality is rapid. Thus, comparative studies on OTU demarcation for datasets generated using a specific high-throughput sequencing platform may not be trivially portable to other data types, and may be 'outdated' within a few years as the field moves on.

For the present analyses, a different approach was chosen, in the tradition of several previous studies [50,57,82]: we generated a global, comprehensive dataset of high-quality, near full-length SSU sequences which were downloaded from NCBI *GenBank*[*] [123] and from the genomes available in the NCBI Reference Sequence Database (*RefSeq*[†], [124]). From these sources, we filtered for sequences that were annotated as 'ribosomal RNA' or 'rRNA' and had a minimum length of 1,000nt. The raw dataset thus contained reads that were either generated by Sanger sequencing [32], or by (curated) assemblies. After additional filtering and pre-processing steps (see next section, 5.1.2), the dataset comprised almost 1 million sequences, which is two to three orders of magnitude larger than test sets used in previous studies. These sequences were sampled from a wide range of environments (see also section 5.2) and represented a comprehensive survey of publicly available full-length SSU data. Indeed, the used dataset resembles in size, scope and sequence processing the curated sets provided by SSU reference databases such as *RDP* [27], *Greengenes* [28] or *SILVA* [125]. These databases are often used as pre-clustered references in *reference-based* OTU-binning approaches, as well as in global-scale ecological studies, as e.g. on microbial interaction networks [126]. Thus, the test dataset used here is highly curated, represents several realistic use cases, and may provide technology-independent insights on general trends in OTU demarcation.

5.1.2 Preprocessing: filtering, alignment and distance calculation

We generated a pseudo-multiple sequence alignment (pseudo-MSA) of the entire dataset from pairwise alignments of sequences to curated covariance models using the alignment software *Infernal* [80,81]. *Infernal* provides very fast and accurate profile-based alignments that take into account the SSU RNA molecule's highly specific secondary structure. All sequences were aligned to reference consensus models of the bacterial and

[*] http://www.ncbi.nlm.nih.gov/genbank/, accessed in April 2012

[†] http://www.ncbi.nlm.nih.gov/RefSeq/, accessed in March 2012

archaeal 16S rRNA molecule and the eukaryotic 18S rRNA molecule as provided in the package *ssu-align*‡ [80]. In a recent study, Wang et al found that for the alignment of 16S sequences, structure-aware approaches such as used by Infernal did not outperform traditional pairwise alignment methods, such as the Needleman-Wunsch algorithm [73]. However, Wang et al used a dataset of relatively short sequences (231nt) from the V2 SSU subregion which exhibits relatively little secondary conformation. To assess alignment quality, they used a *NMI* metric to test accordance with a 'ground truth' dataset; however, this approach is problematic for this particular kind of problem (see also section 3.4.2). Moreover, Schloss has pointed out a series of further limitations in the Wang et al commentary and discussed the use of secondary structure informed alignment methods [84]. In using full-length sequences that have on average a much higher degree of structural information than the V2 region only, we are confident that a structure-aware approach adds accuracy to our alignments.

Sequences were assigned to the three phylogenetic domains of life (archaea, bacteria and eukarya) based on which reference model they aligned to with the highest *Infernal* alignment score; sequences with a negative score for all three models were excluded from analyses altogether (this step also removed remaining non-SSU rRNA sequences). To obtain an alignment of uniform length, comprising the same amount of information per sequence, all sequences were pruned at manually chosen conserved flanking positions (alignment position 142 to 899 for the archaeal model, 107 to 1,408 for bacteria, and 629 to 1,547 for eukarya), yielding three distinct alignments of lengths 757nt, 1,301nt and 918nt, respectively. We filtered for chimeric sequences using *uchime* [41] with a set of reference sequences generated *de novo* from the entire alignments. This way, 18.9%, 19.7% and 9.7% of sequences were identified as chimeric and removed before subsequent analyses. After these pre-processing steps, the dataset used in this study comprised 950,014 sequences (42,024 archaeal, 887,870 bacterial and 20,120 eukaryotic) of which 720,086 or 75.8% were unique (30,962, 673,128 and 15,996, respectively). These sequences each covered (approximately) the entire 16S/18S SSU rRNA molecule.

5.1.3 Extraction of defined 'local' sequence subsets

Two defined, more compact subsets were extracted from the global dataset for further analyses. The *human skin microbiome* (HSM) is a reference dataset of 112,283 sequences (90,620 after quality filtering steps) sampled from 21 distinct human skin sites [51]. The HSM set has been extensively studied, and has previously served to benchmark OTU demarcation [104]. In context of the analyses presented in this thesis, the HSM was used to assess differences between clustering methods with respect to robustness and reproducibility in OTU demarcation and OTU-based ecological data descriptions (section 7.1). Moreover, it was used as an ecologically well-defined model dataset to study the correlation of ecological similarity with SSU sequence similarity (section 7.2), as well as the effects of clustering context (section 7.1). In the latter framework, the HSM represented a 'local' subset of the full 'global' dataset, as it was more closely circumscribed in terms of ecology, taxonomic composition and sequence space.

In the same context, we generated a second 'local' set of *broad ecological range* (BER), comprising 53,999 sequences from 18 studies focusing on distinct environments ([17, 127-139]; see Table 5.1 on the next page).

‡ http://infernal.janelia.org

Reference	Environment Type	Number of Sequences
Eckburg PB et al., 2005 [17]	human intestine	11,831
Grice EA et al., 2008 [127]	human skin	6,209
Shaw AK et al., 2008 [128]	aquatic (ocean)	5,654
Elshahed MS et al., 2008 [129]	soil	2,699
Brulc JM et al., 2009 [130]	bovine rumen	2,918
Alonso-Gutierrez J et al., 2009 [131]	coastal / oil-contaminated	1,606
Cruz-Martínez K et al., 2009 [132]	soil (grassland)	1,034
Walsh DA et al., 2009 [133]	aquatic (ocean, anoxic)	5,810
Sunagawa S et al., 2010 [134]	coral-associated	1,956
Durso LM et al, 2010 [135]	bovine intestine	11,107
Eloe EA et al., 2010 [136]	aquatic (deep sea)	1,271
Perkins SD & Angenent LT, 2010 [137]	metalworking fluid	1,059
Martinson VG et al., 2011 [138]	insect-associated	4,956
Perkins SD et al., 2011 [139]	wastewater sludge	1,144
Unpublished: JJ Wright & SJ Hallam, Microbial structure in the oxygen minimum zone of the Northeast subarctic Pacific Ocean, GenBank Acc No JQ220557 to JQ227301	aquatic (ocean, anoxic)	7,671
Unpublished: VE Olalde-Mathieu et al. Bacterial species richness in an extreme saline-alkaline soil of the former lake Texcoco, GenBank Acc No JN177806 to JN178884	soil (saline-alkaline)	1,060
Unpublished: A Jimenez-Aguilar et al. Bacterial communities in soil under lichen and moss crusts, GenBank Acc No JN023098 to JN024099	soil	1,002
Unpublished: R Zhang & WT Liu, The diversity of bacteria in Tibetan Lake, GenBank Acc No HM126671 to HM130044	aquatic (lake)	3,374

Table 5.1. Sample composition of the artificially generated broad ecological range (BER) local dataset.

5.1.4 Datasets on selected SSU subregions

As current high-throughput sequencing platforms are limited in read length, most SSU-sequencing based studies target specific *hypervariable* subregions of the SSU rRNA gene. To test how the choice of subregion may influence OTU demarcation, we extracted three sets of *short read* sequences from the global alignment of full-length bacterial 16S sequences, notably on subregions *V23*, *V35* and *V6* (see Figure 5.1). While V23 and V35 correspond very closely to sequence subregions used in the *human microbiome project* [34,50], the shorter V6 is a frequent target in Illumina-based studies [53]. By design, the extracted datasets do not correspond to 'real' targeted 454 and Illumina sequencing datasets. Rather, excision from known, full-length reference sequences allows direct comparison of OTU partitions based on full (i.e., *information rich*) and short (*information sparse*) sequences.

	Length	NT Position (Infernal)	NT Position (E. coli ref.)
FL	1301	107-1408	105-1372
V23	429	107-536	105-514
V35	553	378-931	357-906
V6	60	1012-1072	986-1045

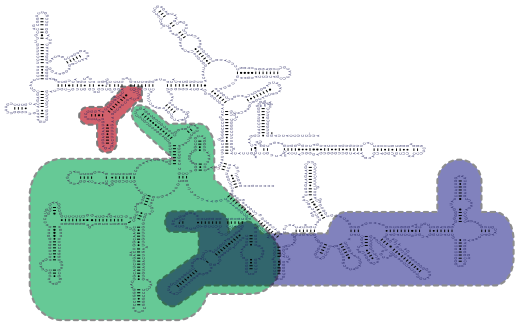


Figure 5.1. Extracting datasets on selected hypervariable SSU rRNA subregions. From the global alignment of full-length bacterial 16S sequences, subregions V23, V35 and V6 were extracted. Subregion sequence length (left column), flanking positions in the Infernal alignment (middle column) and with respect to the E. coli reference 16S sequence (right column) are given. Subregions start and end positions as used by Schloss, 2010 [50]. 16S sequence secondary structure (right panel) modified from an image kindly provided by Harry Noller, University of California, Santa Cruz.

5.2 Contextual data

Both the *GenBank* and *RefSeq* databases provide facilities for submitting rich metadata with each sequence. We harvested this contextual information in several ways to describe ecological properties of the organisms in the SSU sequence dataset. First, we assigned sequences to individual *sampling events*, defined here as a unique combination of submitting authors, publication title and isolation sources; this classified the dataset into 31,519 samples, the largest of which comprised 61,479 sequences, at an average sample size of 30.2 sequences per sample.

Next, we extracted *annotation keywords* for every sample from the publication title, isolation source and additional comments (GenBank annotation field 'note'). We filtered these keywords by removing any terminal letters 's' (to map plural forms) and by requiring that in order to be valid, a keyword had to be used by at least two author teams independently. In addition, as literal taxonomic and geographic annotations carry no ecological information *per se*, we removed all keywords that produced a hit in the NCBI *Taxonomy*[§] database [140] or the *GeoNames*[**] database of geographical place names. Moreover, we used a manually curated list of 1,144 stop words to remove keywords that clearly carried no ecological information (such as the word 'DNA'). In total, these filtering steps reduced the number of annotation keywords by roughly one order of magnitude, yielding 7,202 unique *ecological terms*, at an average frequency of 18.76 samples per term.

---

§ http://www.ncbi.nlm.nih.gov/taxonomy

** http://www.geonames.org, accessed in November 2011

Habitat Type	Habitat Subtype	# of SSU Sequences
anthropogenic	contaminated*	45,990
	wastewater	14,445
	food (fermented)	2,333
	food (dairy)	3,419
	food (other)	32,889
	sterile	3,492
	agricultural	41,348
	other (anthropog.)	29,687
	total (anthropog.)	127,491
host-associated	plant (phyllosphere)	2,197
	plant (rhizosphere)	5,576
	plant (other)	58,095
	skin	342,533
	gastric	63,580
	intestinal	182,003
	oral	3,254
	lung	7,064
	vaginal	803
	blood	1,815
	human host	502,955
	mammalian host	534,766
	insect host	20,166
	animal host (other)	598,106
	total (host-ass.)	643,613
aquatic	marine	64,648
	limnic	21,889
	estuarine	3,836
	littoral	21,274
	pelagic	8,429
	benthic	44,184
	lake	16,459
	stream	6,153
	ice	3,042
	saline	7,806
	other	145,056
	aquatic (total)	230,691
terrestrial (soil)	arctic	2,223
	arid	2,319
	cave	7,382
	forest	3,627
	grassland	11,160
	wetland	9,469
	rock & mineral	14,260
	total (soil)	120,534
thermal	hydrothermal	10,138
	geothermal	12,200
	total (thermal)	19,680
unclassified	total	41,260

Table 5.2: Habitat Classification of 950,014 SSU sequences. Habitat typing was non-exclusive: sequences could be associated with multiple habitat (sub-)types. For example, marine aquatic environments could be further classified as benthic, pelagic or littoral; samples from the 'gastrointestinal tract' were annotated as both 'gastric' and 'intestinal', etc. Note also that host-associated habitats were assigned based on both annotation terms and annotated host information.

**Contaminated habitats were classified into additional subgroups: oil, heavy metal, metal, radioactive and polyaromatic hydrocarbon contamination.*

The vast majority of these terms carried biological information characterizing SSU sequences with respect to their ecological and environmental sampling context. Based on these ecological terms and on host organism annotations (see below), we annotated samples to a list of 53 unique *habitat types* using a manually curated classification scheme (see Table 5.2). Habitat typing was non-exclusive: individual samples could be annotated with different habitat subtypes, e.g. 'aquatic, marine, benthic' or 'forest soil, rhizosphere'.

In a complementary approach, we filtered all keywords for the controlled vocabulary maintained by the Environmental Ontology Project (*EnvO*††) and used the ontology to assign related environmental terms to sequences (e.g., 'lake' and 'pond' were both classified as 'water body'). This procedure yielded 672 unique *EnvO terms* mapping to 16,736 samples – indicating that nearly half of all samples could not be annotated using EnvO. However, having been derived using a dedicated ontology for environmental terms, these keywords carry much ecological information.

We assigned *host taxonomy* to bacterial and archaeal sequences from direct annotations (*GenBank* annotation field 'host') and by inference from annotation keywords (terms matching the NCBI Taxonomy that mapped to higher plants or metazoans were considered to refer to putative host organisms). This yielded 2,422 unique host taxonomies (in total representing 5,850 unique taxonomic categories) for a total of 9,621 samples; the remaining 21,898 samples were considered *not host-associated*. The by far most highly represented host organism was *Homo sapiens* (407,107 sequences mapping to 1,003 samples); in general, animal hosts (572,675 sequences) were much more represented than plant hosts (30,210).

Finally, we inferred geographical sampling site information based on various evidence channels. For 1,928 samples (representing 120,244 sequences), geographical coordinates were extracted from direct annotations (*GenBank* annotation fields 'lat' and 'lon'). For an additional 18,803 samples (representing 253,013 sequences), coordinates were inferred to various levels of spatial resolution based on annotation keywords using the *GeoNames* database of geographical place names. The *GeoNames* database also provides high-resolution elevation information for continental (non-ocean) coordinates. In addition, sampling site elevation (or depth, in the case of water bodies and soil) was inferred from annotation keywords, as well as from the *ETOPO1* 1 arc-minute resolution global relief model [141]. Thus, sampling site latitude, longitude and elevation/depth were inferred for a total of 20,731 samples, comprising 373,257 sequences, which mapped to 4,739 unique coordinates, with almost world-wide coverage.

Taken together, these procedures yielded a highly curated and detailed multi-dimensional ecological description of the sequence dataset. These different channels of ecological information were used to assess the *ecological consistency* of sets of OTUs (see sections 5.6 and 7.2). Moreover, these ecological annotations are currently used in the context of various ongoing projects, e.g. to complement microbial interaction networks with ecological information, and to study microbial biogeography on a global scale.

†† http://environmentontology.org, release date 2011-24-03

5.3 Demarcation of OTUs

Sequences were clustered into OTUs using several established approaches: we executed heuristic methods (*cd-hit, uclust, uparse*) and hierarchical clustering algorithms (HCAs*; average, complete* and *single linkage*). For every applied method, we clustered to different sequence identity thresholds (ranging from 80-100% SSU sequence similarity).

As the tested heuristics implement on-the-fly pairwise sequence (pseudo-)alignment and distance calculation, clustering was performed on unaligned datasets for these methods. We generated OTU sets using *cd-hit*[##] [89,90] in *cdhit-est* mode (which is the default in the *cd-hit-otu* pipeline [49]) on a multicore machine using parallelization and standard parameters. Word lengths of 7, 9 and 11 as parameter for sequence similarity calculation were tested; however, while longer word lengths provided significant speed improvements, the observed differences in both OTU total counts and OTU size distributions were negligible (data not shown) so that only results for word length 11 are discussed. The *uclust*[§§] [92] series of OTU sets was generated using the *uclust* software with the *cluster_fast* option and standard parameters. As *uparse* [69] combines OTU clustering with on-the-fly filtering for chimeric sequences, we used the full, unaligned, non-chimera-filtered sequence dataset for *uparse* runs and subsequently mapped shared sequences between *uparse* partitions and the *uchime*-filtered dataset used for the other clustering methods. Both *uclust* and *uparse* did not cluster the large bacterial dataset to similarity thresholds >99%, due to prohibitive memory requirements of the freely available 32bit versions of these tools.

Hierarchical *average, complete* and *single linkage* clustering were performed using the recently developed in-house software package *hpc-clust* [85]. While *cl* and *sl* partitions were obtained for the whole range of tested similarity thresholds, *al* clustering of the large bacterial dataset was only performed for ≥92% SSU similarity due to high memory requirements of the algorithm. *Hpc-clust* parallelizes the hierarchical clustering task and thus allows to cluster large datasets very rapidly (less than 3h wall time for the present dataset of roughly one million sequences on a 256 core computer cluster), while still computing the entire pairwise distance matrix, avoiding any heuristic shortcuts. Moreover, the software provides the option to use different *alignment distance* calculation functions; however, since the OTU sets generated by different tested distance calculation methods showed only negligible differences in terms of ecological consistency (data not shown), we present only results obtained using the *'one gap'* alignment distance calculator, counting gaps of any length between sequences as single mismatches.

Finally, we attempted to cluster the sequence dataset with the commonly used software tools *mothur* ([60], version 1.27.0, 2012-08-08) and *ESPRIT-Tree* ([72], version 1, 2011-11-15). However, we were unable to process the entire dataset of roughly one million full-length sequences, or even smaller subsets of ≥100k sequences with either of these programs, even when providing excessive computational resources (running on a multicore computer with 1TB RAM); this is most likely due to the computationally expensive calculation of the pairwise SSU sequence distance matrix. However, it has been shown for smaller test sets that *mothur* and *hpc-clust* provide virtually identical partitions for the tested HCAs [85]. Moreover, *ESPRIT* and *ESPRIT-Tree* are slightly heuristic approximations of the *cl* and *al* algorithms, although they rely on pairwise rather than multiple sequence alignments. Thus, any findings reported for the tested HCAs are probably portable to *ESPRIT*, *ESPRIT-Tree* and *mothur*.

[##] http://weizhong-lab.ucsd.edu/cd-hit/, version 4.5.4, Build 2012-08-25

[§§] http://drive5.com/usearch/, version 6.0.307

5.4 OTU-based estimators of microbial 'diversity'

The concept of 'diversity' in the context of ecology appears to be surprisingly fuzzy: in 2010, Tuomisto noted that "the term 'diversity' has been used in at least four conceptually different ways in the ecological literature" [142]. The probably most widely accepted general definition and classification of diversity was introduced by Whittaker [143,144] who described (i) **α**-*diversity* as the mean diversity of types in a local habitat, (ii) **β**-*diversity* as differentiation among these habitats and (iii) **γ**-*diversity* as the global diversity of the entire 'landscape', or scope of study. However, in particular the terms '**α**-diversity' and '**β**-diversity' have since been used by different authors to refer to a range of different phenomena using a wide array of mathematical indices [142,145]. Since an in-depth discussion of ecological definitions of diversity exceeds the scope of this thesis, I will here use a terminology that follows the general conventions in the field of microbial ecology (which diverge from conventions in other research fields). In particular, I will use the term '**α**-diversity' to refer to the *local* diversity of an individual sample, and '**β**-diversity' to describe the similarity / dissimilarity *between* communities (samples). Moreover, for the sake of simplicity I will also introduce definitions in microbial ecology lingo, generally referring to 'OTUs' and 'taxa' rather than 'species' and 'classes', and to 'sequences' in 'samples' rather than 'individuals' in 'habitats'.

While many different measures have been used to assess microbial diversity [146], the focus here will be exclusively on OTU-based (or 'species-based') indices, as these are the most relevant to the work presented in this thesis. In particular, this means that phylogeny-informed diversity estimators, such as *Phylogenetic Diversity* or *UniFrac* [147], are not discussed: conceptually, such indices rely on phylogenetic distance, and OTU clustering with subsequent choice of cluster representatives is usually used merely to reduce data complexity prior to phylogeny inference.

5.4.1 Estimating community richness and evenness (α-diversity)

Many approaches to assessing local community diversity have been proposed; as of 2014, the online documentation of the widely-used *mothur* suite alone lists 20 different **α**-diversity indices[***]. The most basic descriptor for local diversity is taxonomic *richness*, or total species count in the sample. As OTUs are used as proxies for taxa at different levels of resolution, total OTU counts are often used as richness estimates in microbial ecology. Noting that sampling bias may cause an underestimation of richness due to unseen taxa, Chao (1984) proposed an abundance-informed richness estimator which is often referred to as '*Chao1 index*' [148]:

$$S_{Chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)}$$

where S_{obs} is the observed richness (number of OTUs) and n_1 and n_2 are the number of *singleton* (only one representative) and *doubleton* (two representatives) taxa, respectively. In other words, the Chao1 index approximates the number of *unseen* taxa based on the number of observed *rare* taxa. As SSU sequencing datasets often exhibit strong abundance skews towards singleton/doubleton OTUs, the Chao1 index and its more generalized form, the *abundance-based coverage estimator* [149,150], have been widely used in microbial ecology.

[***] http://www.mothur.org/wiki/Calculators

The *Shannon* index takes into account community *evenness* (the taxa abundance distribution) in an entropy-based formulation [151]:

$$H = -\sum_i \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right)$$

where n is the total number of sequences and n_i is the size of class i. In other words, the Shannon index describes the uncertainty when determining the OTU membership of a given sequence in the sample (see also section 5.5.2).

The *Simpson* index, proposed by Simpson in 1949, measures "the degree of concentration or diversity achieved when two individuals of a population are classified into groups" [152]. It has been used in various versions in the ecological literature; the most intuitive form is arguably the *inverse Simpson* index:

$$ISI = \sum_i \frac{n_i(n_i - 1)}{n(n-1)}$$

Thus defined, the inverse Simpson index is the inverse probability that two representatives (sequences) randomly drawn from a sample belong to the same group – high ISI values indicate higher 'diversity'.

All three indices – Chao1, Shannon and inverse Simpson – thus capture different aspects of 'α-diversity', and all three have been widely used in microbial ecology. In the work presented here, we used these indices to characterize the *human skin microbiome* dataset (see section 5.1.3) when clustering sequences according to different methods (section 7.1).

5.4.2 Estimating community similarity (β-diversity)

Similarly to *local* community diversity, many approaches have been proposed to assess the similarity *between* communities ('β-diversity'): as of March 2014, the online documentation for *mothur* lists 36 distinct β-diversity calculators, while the software package *phyloseq* [153] provides as many as 45 options, often with additional parameters. The most basic metric of overlap between groups was first proposed by Jaccard in 1901 [154] who described community similarities for alpine flowers as the fraction of shared species, calculated as ratio of shared species (set intersect) and total species between groups (set union):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In 2004, Chao et al [155] proposed a probabilistic abundance-informed version of the Jaccard index which corrects for unseen taxa:

$$J_{abd}(A,B) = \frac{U_{est} V_{est}}{U_{est} + V_{est} - U_{est} V_{est}}$$

where U_{est} and V_{est} are the (unseen taxa-corrected) estimates of total relative abundances of shared species in groups A (U_{est}) and B (V_{est}), defined as:

$$U_{est} = \sum_{i}^{S_{A,B}} \frac{a_i}{n_A} + \frac{n_B - 1}{n_B} \frac{f_{+1}}{2 f_{+2}} \sum_{i}^{S_{A,B}} \frac{a_i}{n_A} I(b_i = 1)$$

$$V_{est} = \sum_{i}^{S_{A,B}} \frac{b_i}{n_B} + \frac{n_A - 1}{n_A} \frac{f_{1+}}{2 f_{2+}} \sum_{i}^{S_{A,B}} \frac{b_i}{n_B} I(a_i = 1)$$

where $S_{A,B}$ is the number of shared OTUs between groups A and B, a_i is the size of OTU i in A, b_i the size of OTU i in B, n_A and n_B are the total number of sequences in A and B. *I(expression)* is an indicator function, defined as I = 1 if 'expression' is true and I = 0 otherwise. Finally, f_{+1} and f_{+2} are the number of shared OTUs that are singletons and doubletons in partition A, while f_{1+} and f_{2+} are the number of shared OTUs that are singletons and doubletons in partition B. Thus, the number of 'unseen shared taxa' is estimated based on the number of 'observed shared rare taxa' between the partitions. In the above formulation, the Jaccard and abundance-corrected Jaccard indices are defined as community *similarities*, so that $J = 1$ describes perfectly identical communities, while $J = 0$ if no taxa are shared.

The *Sørensen-Dice-Czekanowski* index (*SDC*) [156] is mathematically closely related to the Jaccard index, and various incidence-based and abundance-based formulations have been proposed. In the raw abundance-based version after Chao et al, 2004 [155], it is defined as follows:

$$SDC = \frac{2UV}{U+V} = \frac{2 \sum_{i}^{S_{A,B}} \frac{a_i}{n_A} \sum_{i}^{S_{A,B}} \frac{b_i}{n_B}}{\sum_{i}^{S_{A,B}} \frac{a_i}{n_A} + \sum_{i}^{S_{A,B}} \frac{b_i}{n_B}}$$

where U and V are the sums of relative abundances of individuals in shared taxa in groups A and B. In the above formulation, the *SDC* is defined as an index of community *similarity*; it is closely related to the widely used *Bray-Curtis dissimilarity* index [157].

Another frequently used measure of community similarity is the Morisita-Horn overlap index [158], defined as:

$$MH = \frac{2\sum_{i}^{S} a_i b_i}{\left(\sum_{i}^{S} \frac{a_i^2}{n_A^2} + \sum_{i}^{S} \frac{b_i^2}{n_B^2}\right) n_A^2 n_B^2}$$

where S is the total number of unique taxa between groups, n_A and n_B are the total number of sequences in groups A and B, and a_i and b_i are the absolute frequencies of taxon i in A and B. Values for MH range between 0 (no overlap between communities) and 1 (all taxa are present in both groups in equal abundances).

We used the J_{abd}, SDC and MH indices to assess pairwise community similarities for the samples of the *human skin microbiome* dataset (see section 5.1.3) when clustering sequences according to different methods (section 7.1).

5.5 Assessing partition similarity

When clustering the same set of sequences under different parameters, the resulting data partitions may vary in (i) total cluster counts, (ii) cluster size distributions and (iii) cluster composition. Divergent cluster counts are usually straightforward to interpret: they indicate quantitative overall differences between two partitions (i.e., 'how many clusters are formed?'), but provide no inherent information on qualitative differences (i.e., 'which sequences cluster together?'). However, in the context of OTU clustering, total cluster counts indeed carry biological meaning to some extent, as they are a proxy for taxonomic richness (see section 5.4.1). In consequence, several previous studies have relied on total OTU counts to compare sets of OTUs, often assessing the relative overestimation of diversity with respect to known 'ground truth' partitions ([50,53,66,71,82,83,109]; see section 3.4.1). Similarly, differences in cluster size distributions may indicate quantitative differences between partitions ('how large are the clusters that are being formed?'). In the context of microbial communities, differentially skewed OTU size distributions may directly correspond to differential estimates of community evenness, while also more generally influencing abundance-informed measures of α- and β-diversity (see section 5.4).

The quantification of differences in cluster *composition*, in contrast, is non-trivial and continues to be an open problem in the fields of statistics and machine learning [120,159]. An in-depth discussion of the many approaches to assessing set similarity exceeds the scope of this thesis; rather, I will focus on those measures that were used in this work, which fall into the categories of (i) *pair counting-based* and (ii) *information theoretic-based* indices. Moreover, I will side-step rigorous mathematical formalism by referring to the specific problem of OTU clustering only, discussing clusters of 'sequences' and 'OTUs' rather than using the more abstract terminology of clustered 'nodes' or 'data points', while also using the terms 'partition', 'clustering' and 'OTU set' interchangeably.

5.5.1 Pair counting-based indices

As indicated already by name, pair counting-based indices quantify the similarity between two partitions by counting individual pairs of sequences as either *concordant* or *discordant*. A pair of sequences is concordant across partitions if it either clusters together in both partitions (*'agree to agree'*; Fig. 5.3, left) or does not cluster together in either partition (*'agree to disagree'*; Figure 5.3, middle). In contrast, discordant sequence pairs cluster together in one partition, but into different OTUs in the other.

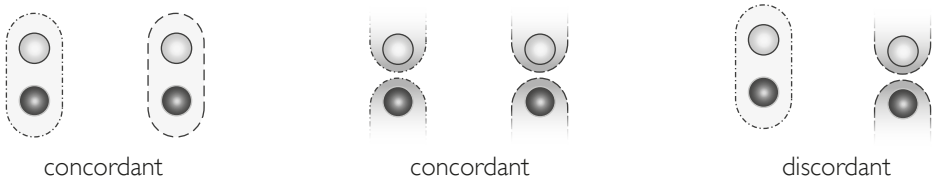


Figure 5.3: concordant and discordant pairs of sequences.

The *Rand Index* [160] of partition similarity weighs counts of concordant and discordant pairs of sequences as follows:

$$RI = N_{concordant} / \binom{n}{2}$$

where n is the total number of sequences and $N_{concordant}$ is the number of concordant pairs. In other words, the Rand Index is the ratio of concordant pairs per total pairs. Based on the observation that the Rand Index does not take a constant expected value between random partitions, Hubert and Arabie proposed an adjusted form which corrects for chance based on a hypergeometric randomness model [161]. The *Adjusted Rand Index* (*ARI*) is calculated as follows:

$$ARI = \frac{Index - Expected\_Index}{Max\_Index - Expected\_Index} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where a_i is the size of OTU i in partition A, b_j is the size of OTU j in partition B and n_{ij} is the number of sequences clustering into OTU i in partition A and OTU j in partition B (i.e., the i,j-th entry in the contingency table between partitions). *ARI* values range between -1 (complete discordancy – sequences grouping together in A never group together in B) to 1 (perfectly identical partitions); *ARI = 0* indicates random similarity as expected based on cluster size distributions. We used the *ARI* to compare partitions in a range of different experiments (sections 6.1, 6.2 and 7.1).

Arguably, one inherent drawback of pair counting-based measures is the dominance of large clusters. Since the number of pairwise comparisons scales quadratically with sequence count, large clusters will contribute disproportionally more to the similarity / dissimilarity signal than smaller clusters. In an extreme case, merging two large clusters X_A and Y_A in partition A into one cluster Z_B in partition B will provide a large number of discordant sequence pairs, although both partitions remain highly similar from a set point of view.

5.5.2 Information theoretic-based indices

More recently, information theoretic-based indices have received increasing attention in the clustering literature, not least due to their strong theoretical background [120]. Consider a partition A of i clusters of sizes a_i. The *entropy* of partition A quantifies the uncertainty when determining a given sequence's cluster membership in A; it is calculated as follows:

$$H(A) = -\sum_i \frac{a_i}{n} \log\left(\frac{a_i}{n}\right)$$

H(A) = 0 indicates the 'singleton partition', i.e. a partition with only one cluster in which there is no uncertainty about cluster membership. The information overlap between two partitions A and B can be expressed based on these partitions' entropies, as the *Mutual Information* (MI):

$$I(A,B) = \sum_i \sum_j n_{i,j} \log\left(\frac{n_{i,j} n}{a_i b_j}\right)$$

In other words, *I(A,B)* quantifies the mutual dependence between partitions A and B and "measures how much knowing one of these [partitions] reduces our uncertainty about the other" [120]. As I(A,B) is upper bounded by the entropies H(A) and H(B), several mathematically related or even equivalent normalizations have been proposed, such as the *Variation of Information* (*VI*) by Meilă [162] or different versions of the *Normalized Mutual Information* (*NMI*). As defined by Fred & Jain, 2003 [163], *NMI* is calculated as follows:

$$NMI = \frac{-2I(A,B)}{H(A)+H(B)} = \frac{-2\sum_i\sum_j n_{i,j}\log\left(\frac{n_{i,j}n}{a_i b_j}\right)}{\sum_i a_i \log\left(\frac{a_i}{n}\right)+\sum_j b_j \log\left(\frac{b_j}{n}\right)}$$

NMI values range between 0 (no shared information between partitions) to 1 (perfectly identical partitions). In the context of OTU demarcation, both *NMI* and *VI* have previously been used to test the agreement of OTU sets with differently defined taxonomic ground truth partitions [58,66,72,82,100]; see also section 3.4.2.

Noting that variations in cluster counts cause systematically shifting *NMI* baseline values, Vinh et al. proposed the *Adjusted Mutual Information* (*AMI*) measure which uses a hypergeometric permutation model to correct for these effects [120]:

$$AMI = \frac{I(A,B)-E\{I(M)\,|\,a,b\}}{\sqrt{H(A)H(B)}-E\{I(M)\,|\,a,b\}}$$

where $E\{I(M)|a,b\}$ is the expected average Mutual Information for all theoretically possible contingency tables with marginals *a* and *b*; in other words, $E\{I(M)|a,b\}$ is the expected Mutual Information for the observed distributions of cluster sizes in partitions A and B. It is defined as

$$E\{I(M)\,|\,a,b\} = \sum_i\sum_j\sum_{n_{i,j}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{i,j}}{n}\log\left(\frac{n_{i,j}n}{a_i b_j}\right)\frac{a_i!b_j!(n-a_i)!(n-b_j)!}{n!n_{i,j}!(a_i-n_{i,j})!(b_j-n_{i,j})!(n-a_i-b_j+n_{i,j})!}$$

Vinh et al could show in simulation studies that *AMI* values do not suffer from a systematically shifting baseline with shifting cluster counts. Similarly to *ARI*, values for *AMI* range between [-1, 1]; *AMI = 1* describes perfectly identical partitions, *AMI = 0* indicates 'random' shared information as expected by chance for two partitions of the given cluster size distributions. We used both *NMI* and *AMI* to assess partition similarity across clustering methods, and for varying clustering parameters (see sections 6.1, 6.2 and 7.1).

5.6 The *Ecological Consistency Score* as a measure of OTU set ecological consistency

We developed an *Ecological Consistency Score (ECS)* to assess the ecological consistency of entire sets of sequence clusters with respect to different ecological signals. An in-depth theoretical and empirical motivation for the ECS is given in section 7.2.

Consider an individual OTU i clustering n_i sequences from different sampling events. Each sequence is annotated according to different ecological signals characterizing the environment from which it was sampled, such as e.g. ecological terms or host organism taxonomy (see section 5.2). We consider OTU i to be 'ecologically consistent' if it is enriched in sequences that share similar ecological affiliations. We calculated the likelihood $L_{i,j}$ of observing any biological feature j (e.g., an ecological term such as 'soil', 'skin' or 'ocean') with a global background frequency p_j in the entire dataset exactly ki_j times in an OTU i of size n_i using a binomial model:

$$L_{i,j} = \begin{pmatrix} n_i \\ k_{i,j} \end{pmatrix} p_j^{\,k_{i,j}} \left(1 - p_j\right)^{n_i - k_{i,j}}$$

For example, observing 5 sequences annotated with the ecological term 'skin' (background frequency in the global dataset of 30.0%) in an OTU containing 15 sequences has a likelihood of 0.206, but observing the much less frequent term 'hydrothermal' (background frequency ~0.9%) exactly 5 times in the same OTU is much less likely ($L_{15,hydrothermal} = 1.6*10^{-7}$). Similarly, not observing a frequent term such as 'skin' in the same OTU has a rather low likelihood ($L_{15,skin} = 0.005$). Thus, the presence of 5 sequences annotated as 'hydrothermal' in an OTU of size 15 is an *enrichment of ecologically similar organisms*, while the absence of a frequent term such as 'skin' in the same OTU is a *negative enrichment*.

While $L_{i,j}$ describes the ecological consistency of an individual OTU, what is the likelihood of the enrichment of *all* ecological features across *all* sequence clusters in the dataset? We computed this as the summed log-likelihood LL_{set} over all $L_{i,j}$:

$$LL_{set} = \sum_i \sum_j \log(L_{i,j})$$

High absolute values of LL_{set} indicate that enrichments of ecological features in OTUs across the entire partition are non-random. However, the absolute value of LL_{set} is influenced by total OTU count (as the number of summands i) and OTU size distribution (as n_i in the binomial coefficient). Thus, in order to compare biological consistency between OTU sets, we used an empirical approach to control for these effects. For any given OTU set, we generated 1,000 randomized sets with identical OTU size distribution, but shuffled sequence-to-OTU mapping and computed the summed log-likelihood LL_{rand} for each of these sets. This generated near-Gaussian distributions of randomized set log-likelihoods LL_{rand}. From this, we calculated the *ecological consistency score* of the observed OTU set as standard Z score:

$$ECS = -\frac{LL_{set} - \mu_{rand}}{\sigma_{rand}}$$

where μ_{rand} is the average value of LL_{rand} and σ_{rand} is the standard deviation. Thus, *ECS* values indicate by how many standard deviations the enrichment of ecological features in an observed OTU set is removed from a randomized background. In other words, the *ECS* indicates how consistent a given set of OTUs is with respect to an ecological signal, such as the distribution of ecological terms. *ECS* values are independent of both OTU

size distribution effects and total number of OTUs in the set and provides a measure that is comparable between OTU sets.

We used an empirical jackknifing approach to assess *ECS* variability. For a given data point, 1,000 likelihoods of randomized sets (LL_{rand}, see above) were calculated, from which we recalculated *ECS* values based on 1,000 subsamples of 100 LL_{rand} values. This provided a jackknifed estimate of *ECS* mean values and standard deviations. Based on these *ECS* distributions, a Student's t-test was used to test for significance in *ECS* differences between methods.

6. Major Results and Discussion

6.1 The choice of clustering method biases biological data interpretation

The advent of high-throughput sequencing technology has been a mixed blessing for the research field of microbial ecology. The ability to generate very deep sequencing datasets for virtually any environment from which microbial DNA can be isolated has been a great asset, as it has enabled studies at previously unattainable scopes: over the past few years, very large targeted datasets for many environments have emerged which have considerably advanced our understanding of microbial diversity. However, as in many other fields, the development of computational analysis tools has struggled to keep up with the rapid accumulation of available sequence data. The demarcation of OTUs as clusters of marker gene sequence similarity is one attempt to conquer this increased data complexity: as analyses at the level of individual sequences are usually too resource-demanding for current-generation computers, clustering by sequence similarity is supposed to reduce the problem to computationally accessible scales. However, it has to be borne in mind that marker gene-based OTUs are, technically speaking, *proxies of proxies for proxies* – a marker gene is a proxy for taxonomic identity, a sequence similarity cluster approximates 'true' bacterial taxa, and the predictive value of rivaling 'bacterial species' concepts with respect to 'true' microbial lineages remains highly controversial.

The challenge of providing efficient, scalable and accurate OTU demarcation from complex sequencing datasets has triggered the development of a myriad of methods and software tools. Generally, OTU demarcation pipelines provide flexibility in parameter choices at many different levels; in this thesis, the explicit focus has been on sequence clustering methods. Generally, different algorithms implement divergent, and sometimes mutually contradictory, basic assumptions on the fundamental organization of microbial diversity – see for example the discussion on *inclusive* and *exclusive* clustering in section 3.2.5.1. Nevertheless, OTUs are often used as synonymous surrogates for 'species': they provide the backbone for further ecological data description and quantitative biological interpretation. However, while a substantial body of literature on unidimensional benchmarks of clustering methods against different concepts of 'optimality' is available, surprisingly little is known about how similarly or dissimilarly they partition datasets. In other words, a truly quantitative understanding of the variability introduced by sequence clustering has been lacking.

In the work presented in this thesis, we have quantified this variability for six representative clustering methods: hierarchical *average, complete* and *single* linkage clustering, as well as heuristic *cd-hit*, *uclust* and *uparse* clustering. Manuscript 7.1 discusses a series of experiments which assessed partition similarities at different levels. We confirmed previous observations that clustering methods provide markedly divergent total OTU counts, and we complemented this finding by showing how these divergences translate to the level OTU size distributions. More strikingly, we observed characteristic differences between methods in cluster *composition*. Figure 6.1 provides a palpable illustration of these differences for a single datapoint, clustering the *human skin microbiome* (HSM) dataset to 97% nominal sequence similarity. While we observed fluctuating cluster membership of sequences between partitions, there were very few cases of 'truly' discordant clustering (sequences that traversed OTU 'boundaries'). Rather, most of the observed variation was introduced by differential 'lumping' and 'splitting' of defined sets of sequences, with *sl* usually providing the most comprehensive (*inclusive*) clusters, while *uclust* and *cl* provided strongest levels of sub-partitioning.

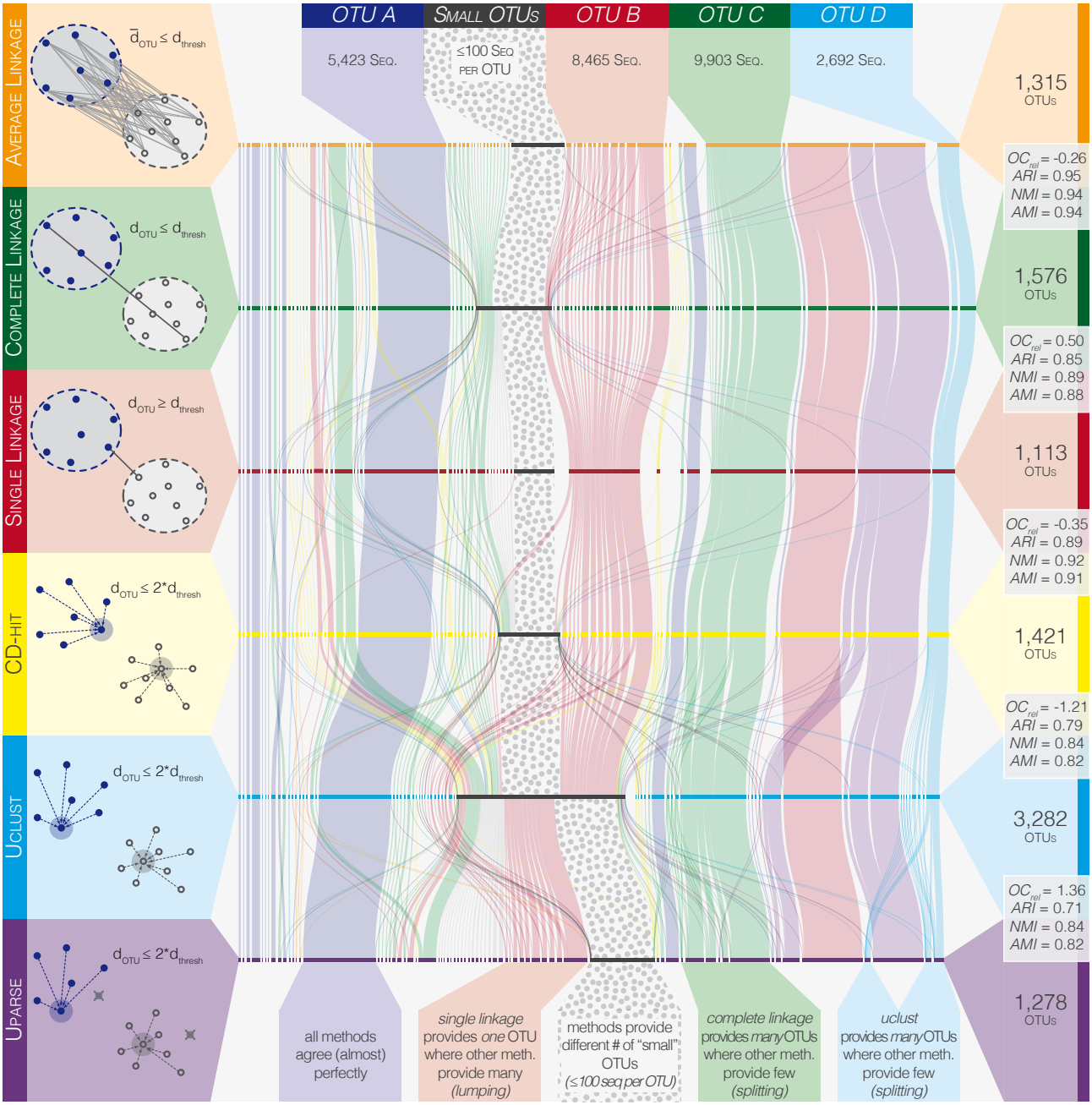


Figure 6.1. Differences in cluster composition when clustering the HSM dataset to 97% nominal sequence similarity according to different methods (toy models of algorithms in left panel). OTU sets (horizontal bars) showed characteristic variations in sequence composition (vertical bands) between methods, as quantified using different measures of partition similarity (right panel). Figure reprinted with detailed discussion in manuscript 7.1.

For the same exemplary datapoint, we quantified the influence of the choice of clustering method on ecological data description, based on different estimators of **α**- and **β**-diversity for the 21 skin habitats represented in the HSM dataset. We found that some methods (especially *uclust* and *sl*) provided systematically shifted diversity estimates, while others (*al*, *cd-hit* and *cl*) provided highly similar descriptions of the data. Moreover, clustering methods also ranked habitats differently by local diversity and pairwise community similarity. Thus, bias in ecological data description was introduced on two conceptual levels: absolute diversity estimates were systematically shifted between methods, while trends between habitats were also influenced by method choice. We could show that these trends and shifts in diversity estimates between methods correlated with differences in cluster composition (as expressed in *AMI*, *NMI* and *ARI* values).

6.2 Clustering methods are differentially robust to changing parameters

We generalized our observations on differential cluster composition between methods to a global dataset of 887,870 bacterial 16S sequences, and to a wide range of clustering similarity thresholds. Figure 6.2 shows pairwise partition similarities, expressed as *Adjusted Mutual Information* (*AMI*), between methods across thresholds. We found that in general, maximum partition similarities between methods were *off-diagonal* – that is, comparisons at the same nominal similarity threshold (e.g., '97%') were generally suboptimal. Moreover, while some methods provided highly similar partitions across wide threshold ranges (AMI ≥ 0.9 for *al, cd-hit* and *cl*), others diverged distinctly in similarity to other methods (e.g., *uclust*) or were more sensitive to changing thresholds (*sl*). These observations are best summed up as *differential reproducibility*: while *al*, *cl* and *cd-hit* provided generally similar partitions of the data, *sl, uclust* and *uparse* provided more dissimilar behavior to all other methods. We quantified this observation by assessing pairwise similarity between methods as correlation of similarities to all other methods (see manuscript 7.1 for a detailed discussion). Moreover, we confirmed these trends for the smaller HSM dataset, as well as for other measures of partition similarity (*NMI, ARI*).

We also assessed reproducibility of clustering methods against 'themselves', i.e. when partitioning the same set of sequences twice, but in randomized input order (plots on diagonal in Figure 6.2). When clustering twice to the exact same nominal threshold, methods provided identical (*al, sl, cd-hit, uparse*) or nearly identical (*cl, uclust*) partitions; it has to be noted, however, that for the heuristic methods these high levels of replicability were probably due to internal sequence sorting, rather than deterministic algorithm behavior. However, when slight changes in clustering threshold were introduced (a step-size of 0.2% similarity corresponded to ~2.6 mutations across the entire length of 1,301 alignment columns), partition similarities dropped sharply for some methods (*uclust, uparse*), while others were more robust (*sl, cd-hit, al*) or indeed highly robust across wide threshold ranges (*cl*). Similar trends in robustness to slightly changing thresholds were observed for comparisons *between* methods. They were also confirmed for the HSM dataset, and for *NMI* and *ARI*.

Manuscript 7.1 also discusses the robustness of methods to other parameters, notably to clustering context and the choice of sequenced SSU gene subregion. By 'clustering context' we refer to the scope of the sequence space: do clustering methods partition a 'local' dataset of defined taxonomic and ecological scope differently when the sequence space is enriched, or thinned out (i.e., more sparse)? In other words, is e.g. the 'local' HSM dataset clustered reproducibly in the presence and absence of the 'global' sequence space provided by the full, comprehensive SSU dataset? We found that context does indeed influence clustering outcome, but that different methods are differentially robust to context effects. In particular, *al,* and to a lesser extent *cl,* provided highly similar partitions of local datasets regardless of context, while other methods, and in particular *uclust*, were more strongly affected (see Figure 5 in manuscript 7.1). Similarly, methods were differentially robust to the choice of SSU sequence subregion, i.e. when comparing partitions of full-length sequences to partitions based on subregions V23, V35 or V6. While *al, cl* and to a lesser extent *cd-hit* were highly robust, *sl* and in particular *uclust* diverged markedly in cluster composition between full-length and subregion partitions (see Figure 6 in manuscript 7.1).

Our observations in manuscript 7.1 can be summarized in three general statements: (i) clustering methods partitioned data differently, and differences in cluster composition could be quantified across thresholds; (ii) methods provided differential levels of *reproducibility*, both in terms of 'within-method' reproducibility (*replicability*) and 'across-method' reproducibility; (iii) methods were differentially *robust* to changing clustering parameters, such as similarity threshold, clustering context and choice of SSU subregion, and trends in robustness between methods were consistent across tests.

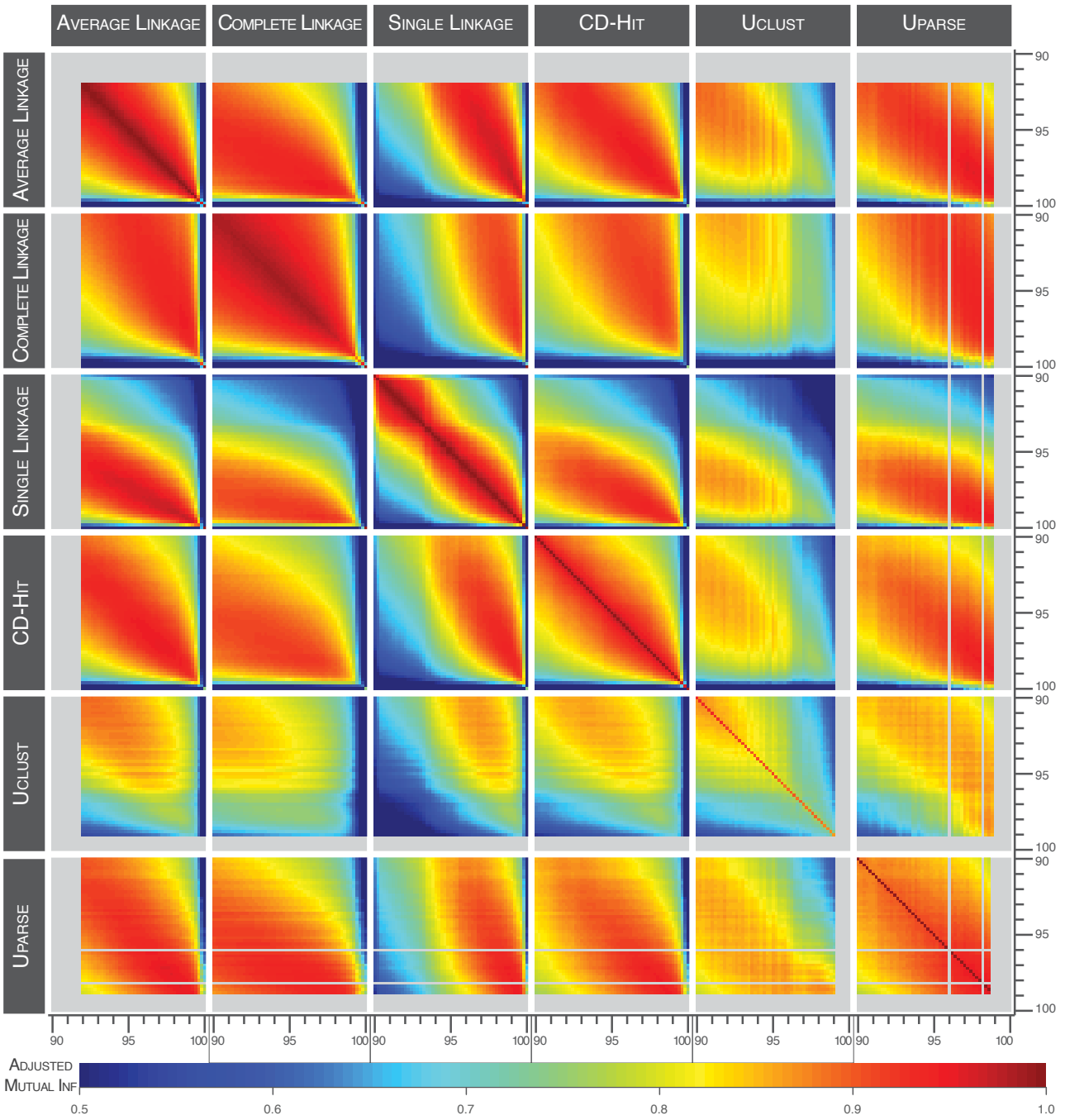


Figure 6.2. Clustering methods are differentially robust to changing similarity thresholds. A global dataset of 887,870 bacterial 16S sequences was clustered according to six different methods (rows and columns) to thresholds of 90-100%, in steps of 0.2%. Partition similarity across methods and across thresholds was assessed as Adjusted Mutual Information (AMI). Figure reprinted with detailed discussion in manuscript 7.1.

These observations are not, in themselves, indicative of clustering 'quality' – we compared methods to each other, with no notion of 'true' or 'false' clusterings. Nevertheless, our findings may be relevant and informative to practitioners in the field, as they reveal several important trends that previous studies have missed. First, our results indicate that biological findings may be comparable between studies for *al*, *cl* and *cd-hit* clustering, albeit within limits. In other words, OTU-based ecological descriptions may be portable between datasets for these methods. In contrast, *sl*, *uparse* and in particular *uclust* diverged so markedly from other methods that findings based on these methods may not be trivially transferable between studies in general.

Second, small changes in clustering threshold may have a strong impact on biological interpretation for some methods, in particular for *uclust* and *uparse*. Considering that the tested step size of 0.2% is much smaller than differences in clustering thresholds between studies usually are, a meaningful comparison of *uclust*- and *uparse*-based results across studies appears to be virtually impossible. We also note that slight changes in sequence length may likely have similar effects, as the 0.2% step size corresponded to an information gain or loss of ~2.6 nt across the 1,301 alignment columns. In contrast, results for *cl, al, cd-hit* and *sl* may be generally more comparable across thresholds.

Third, clustering context may differentially impact different methods. We found that a rich, 'global' sequence space led to significantly different partitions of 'local' datasets for some methods (most strikingly, *uclust*), while others (*al, cl*) were surprisingly robust. In practice, the taxonomic composition of an environment of interest is not known *a priori*, and sequence space 'richness' cannot be trivially assessed (unless pairwise sequence distances are calculated, which is computationally expensive). For some methods, OTU demarcation may therefore be biased by the taxonomic complexity of the community under study, an effect that cannot be trivially corrected for between different habitats. To our knowledge, effects of clustering context have previously not been described or characterized.

Fourth, the V23 and V35 SSU subregions may provide decent approximations of full-length sequences under some clustering regimes (*al, cl* and to a lesser extent *cd-hit*), while other methods were highly sensitive to the choice of subregion. Several studies have described effects of subregion choice on total OTU counts [50], relative dominance of singleton OTUs [57] and compliance with taxonomic ground truth [58], but to our knowledge, an objective assessment of differential effects on clustering methods has previously been lacking. Indeed, pertaining to wide ranges of clustering thresholds, our findings indicate a generally higher robustness of (some) clustering methods than expected based on the reports by Schloss [50], Schloss & Westcott [57] and Sun, Cai et al [58]. As targeted high-throughput sequencing of SSU subregions continues to gain importance in microbial ecology, our findings may inform study design choices, while also facilitating an informed assessment of previously published findings. For example, we note that *uclust* was the backbone of one of the two complementary SSU analysis pipelines for the *human microbiome project* (HMP) which relied on 454 sequencing of the V13 and V35 subregions.

Fifth, we found that total OTU counts are generally poor indicators of partition similarity. We observed that e.g. *cl* provided partitions which varied markedly in total OTU counts across clustering thresholds, but were highly robust in terms of cluster composition. Similarly, comparing total OTU counts between methods did not reveal similarities in cluster composition, which were usually higher than expected based on relative OTU counts alone. Thus, previous benchmarks which relied on total OTU counts may have missed important trends in partition similarity. In particular, although *cl* clustering has often been reported to 'over-partition' data (i.e., to provide 'too many' OTUs), we found that it was highly robust in cluster composition to all tested parameters.

Sixth, we noted that the different tested heuristics approximated hierarchical clustering differentially well. While *cd-hit* was overall remarkably similar to hierarchical *al* and *cl* clustering, *uparse* and in particular *uclust* diverged markedly across all tests. For *uparse*, this behavior may in part be due to on-the-fly chimera filtering which introduced additional flexibility against the *uchime*-based filtering implemented for all other methods. Differences between *cd-hit* and *uclust* / *uparse* may also in part be due differences in sequence distance calculation. Nevertheless, *cd-hit* approximated hierarchical methods better, and also provided higher robustness and reproducibility than *uclust* and *uparse* in all other tests.

Lastly, we note that the observed levels in robustness and reproducibility were generally higher than expected based on previous reports. In particular, observing similarities across varying clustering thresholds (in 'two dimensions', as shown in Figure 6.2) revealed off-diagonal maxima in partition similarity that unidimensional benchmarks, by design, have failed to capture. Moreover, similarity between methods was robust to wide threshold ranges in some cases (e.g., *cl* vs *al* and *cd-hit*), and partitions often varied much more in relative OTU counts than in cluster composition. Thus, depending on parameters, OTU demarcation for some methods may be comparable across studies, albeit within limits.

6.3 OTUs are generally, though not perfectly, ecologically consistent

The use of OTUs to study microbial communities is fundamentally pragmatic. In theory, 'optimal' descriptions of microbial diversity would rely on fundamental units of diversity that comply with a unifying concept of prokaryotic speciation, thus representing 'true' lineages. However, a commonly accepted bacterial species concept remains elusive to the point of contesting the very existence of bacterial 'species' as such [9,113,164,165] – and OTUs provide a phenomenological approach to the problem of organizing microbial diversity in practice. Defined as clusters of sequence similarity with respect to specific marker genes, OTUs are theory-agnostic, conceptually straightforward and literally '*operational*' fundamental diversity units. However, the question of how well they correspond to 'true' bacterial lineages has been a matter of recent debate. The issue is additionally complicated by the lack of consensus on how 'true' bacterial lineages are defined, or how they can be identified from (small-scale) reference sets of marker gene sequences.

One approach to delimiting 'true' bacterial lineages that has recently received increasing attention is the *ecotype* model of bacterial speciation [105,166]. Ecotypes are defined as ecologically coherent groups of organisms whose diversity is confined by a cohesive genetic force, and thus in principle they reconcile ecological diversity units with evolutionary theory. However, although several dedicated algorithms have been developed to demarcate ecotypes from marker gene sequence datasets [103,104], the significant computational overhead associated with ecotype simulation is prohibitive of their large-scale applicability to real world problems. Moreover, it has been noted repeatedly that recognized diversity clusters within several microbial clades conflict with ecotype theory (e.g., [165,167]).

Another approach to identifying 'optimal' fundamental diversity units has been to benchmark against taxonomic 'ground truth'. In this, the underlying assumption is that shared taxonomy implies phylogenetic and ecological consistency. Indeed, these are two frequently cited criteria for 'good' (i.e., theory-compliant) units of diversity: they should reflect *phylogeny* (by representing *monophyletic* groups of organisms) and *ecology*, since ecological differentiation has been postulated as an important driver of bacterial speciation [9,103,106,121,166,168,169]. But how well do OTUs comply with these criteria?

In the work discussed in manuscript 7.2, we assessed the *ecological* consistency of OTUs; an assessment of their *phylogenetic* consistency has been provided elsewhere and exceeds the scope of this thesis. We clustered a global dataset of 950,014 bacterial, archaeal and eukaryal SSU sequences to varying thresholds according to five different methods (*al, cl, sl, cd-hit* and *uclust*) and assessed OTU ecological consistency based on sequence annotations. Figure 6.3 provides a 'snapshot' of selected OTUs for one datapoint (clustering to 97% sequence similarity), breaking down individual clusters by annotated habitats.

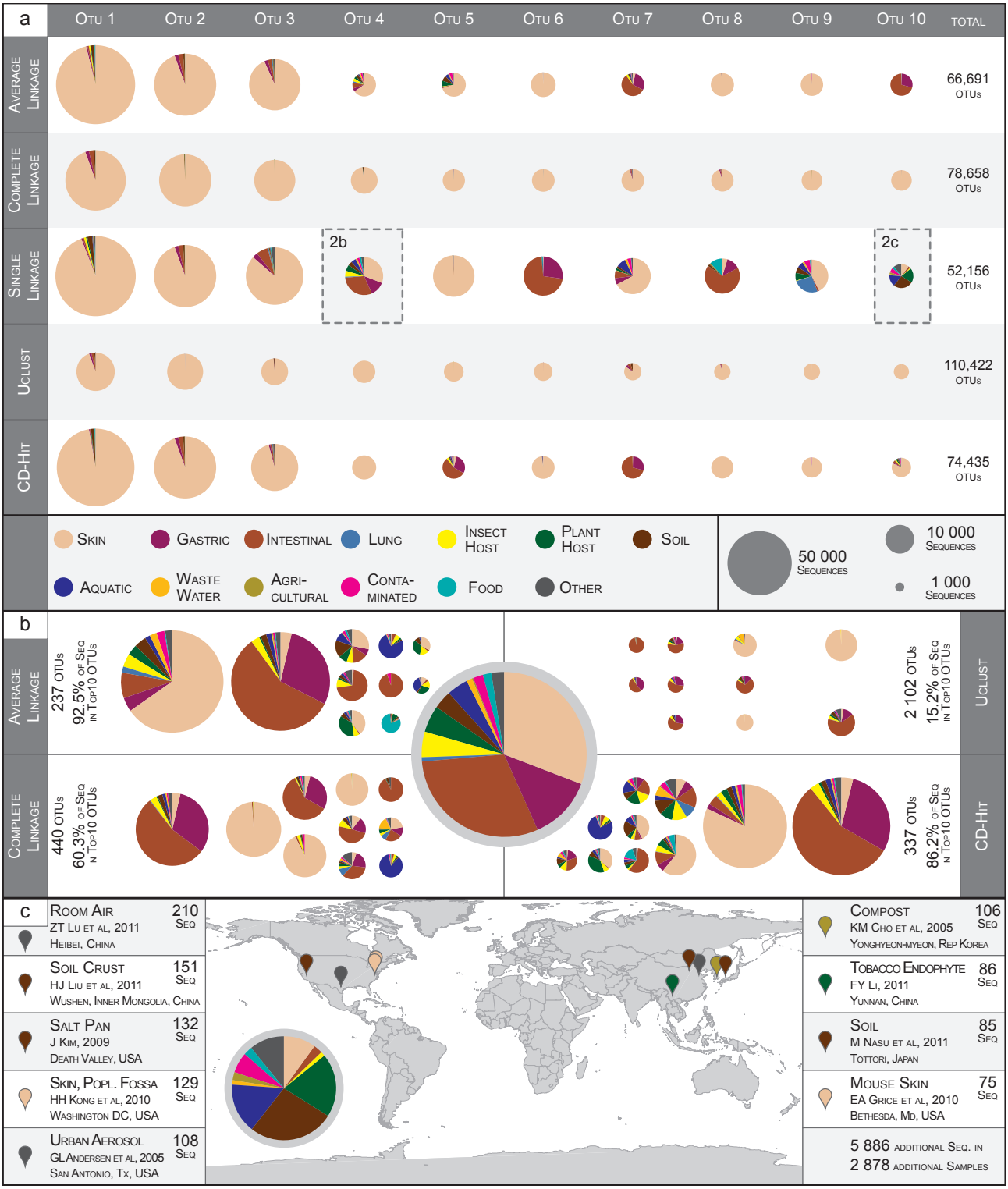


Figure 6.3. Ecological consistency of selected OTUs when clustering a global dataset of 950,014 SSU sequences to 97% sequence similarity according to different methods. (a) Breakdown of the 10 largest OTU per method by habitat annotation. (b) The 17,462 sequences of 'sl OTU 4' were clustered differently, and generally to higher ecological homogeneity, by the other methods. (c) Nine of the largest sampling events contributing to 'sl OTU 10'; this OTU encompassed sequences from very distinct environments. Figure reprinted with detailed discussion in manuscript 7.2.

We observed differential levels of broad-scale ecological consistency for the largest OTUs per clustering method at this selected datapoint. While the overall dominating habitat annotation in the dataset, *skin* (annotated to 30.0% of all sequences) also dominated most of the shown OTUs, there were clear differences in habitat annotations between methods. For *cl* and *uclust*, almost all shown OTUs (except '*uclust* OTU 7') encompassed ≥95% sequences sampled from skin, and almost all remaining sequences were annotated as

'gastric' and 'intestinal'. In contrast, several *sl* OTUs clustered sequences from very diverse environments, with the dominant habitat representing as little as 26.6% of sequences in *sl* OTU 10. Figure 6.3b and 6.3c provide a closer look at selected ecologically inconsistent *sl* OTUs, which contained sequences that were clustered to ecologically more homogeneous OTUs by other methods (6.3b) or were shown to encompass sequences from very distinct environments (6.3c). Generally, this ecological inconsistency of individual *sl* OTUs is likely due to the *inclusive* clustering regime, which may lump sequences together that share below-threshold similarity. For example, although clustering to a nominal threshold of 97% sequence similarity, the mean pairwise similarity between sequences in *sl* OTU 10 (6.3c) was 95.2%, with individual pairs of sequences sharing as low as 86% similarity.

These observations on selected OTUs at an individual datapoint were intentionally anecdotal; moreover, they pertained to very broad ecological scales – from a microbial ecologist's point of view, habitats broadly described as 'aquatic' or 'soil' are arguably sub-structured into a large diversity of ecologically distinct environments. To generalize our observations, we developed an *Ecological Consistency Score* (*ECS*) which quantifies consistency of entire dataset partitions with respect to different ecological features. Intuitively, the *ECS* describes how non-random the enrichment of ecological features in individual clusters is across entire sets of OTUs. Figure 6.4 shows *ECS* values of OTU sets with respect to different ecological features when clustering different datasets to a wide range of similarity thresholds (80-99% sequence similarity). The tested ecological 'features' included *ecological terms* (which provided more fine-scale descriptions than the broad-scale habitat annotations discussed above; 6.4a-c), *EnvO terms* (based on the curated EnvO ontology; 6.4d), *sampling site information* (6.4e) and *host taxonomy* (6.4f); see section 5.2 for a description of these signals. We found that the trends observed in Figure 6.3 for individual OTUs were generally confirmed on a global scale, and across tested thresholds. While at high clustering stringencies (i.e., high OTU counts) different methods provided similar levels of OTU set ecological consistency across tests, pronounced differences were observed with increasing levels of clustering (lower thresholds, fewer OTUs). Trends between methods were robust across tests (*ECS* from highest to lowest): *cl, uclust, all/cd-hit* and *sl*. These differences were statistically significant over wide cutoff ranges ($p < 0.01$, t-test on jackknifed ECS estimates; see section 5.6). The marked drop in *ECS* for *sl* towards lower OTU counts was likely due to incremental lumping of highly dissimilar sequences (see discussion on '*sl* OTU 10' above). Similar effects for *cd-hit* are less straightforward to interpret, but may likely be associated with this heuristic's distance calculation at comparatively low thresholds ($\leq$90%).

We controlled for different factors that could in theory have driven the observed *ECS* trends. By using total OTU counts rather than nominal clustering threshold as the independent variable (i.e., x-axis), we controlled for differential effects of 'over-splitting' or 'over-lumping' relative to other methods at a common nominal threshold. In other words, comparing OTU sets of similar sizes (number of clusters) may to some extent correct for the effects of different concepts of 'clustering distance' (e.g. *inclusive* and *exclusive* regimes) between methods. Moreover, the *ECS* was designed to mathematically correct for OTU count effects, as it takes into account both 'positive' and 'negative' enrichment of ecological features, and implements a conservative empirical randomization scheme. Indeed, trends in *ECS* values between methods did not reflect trends in relative OTU counts; in particular, *uclust* generally provided the highest numbers of OTUs, but not the highest levels of ecological consistency.

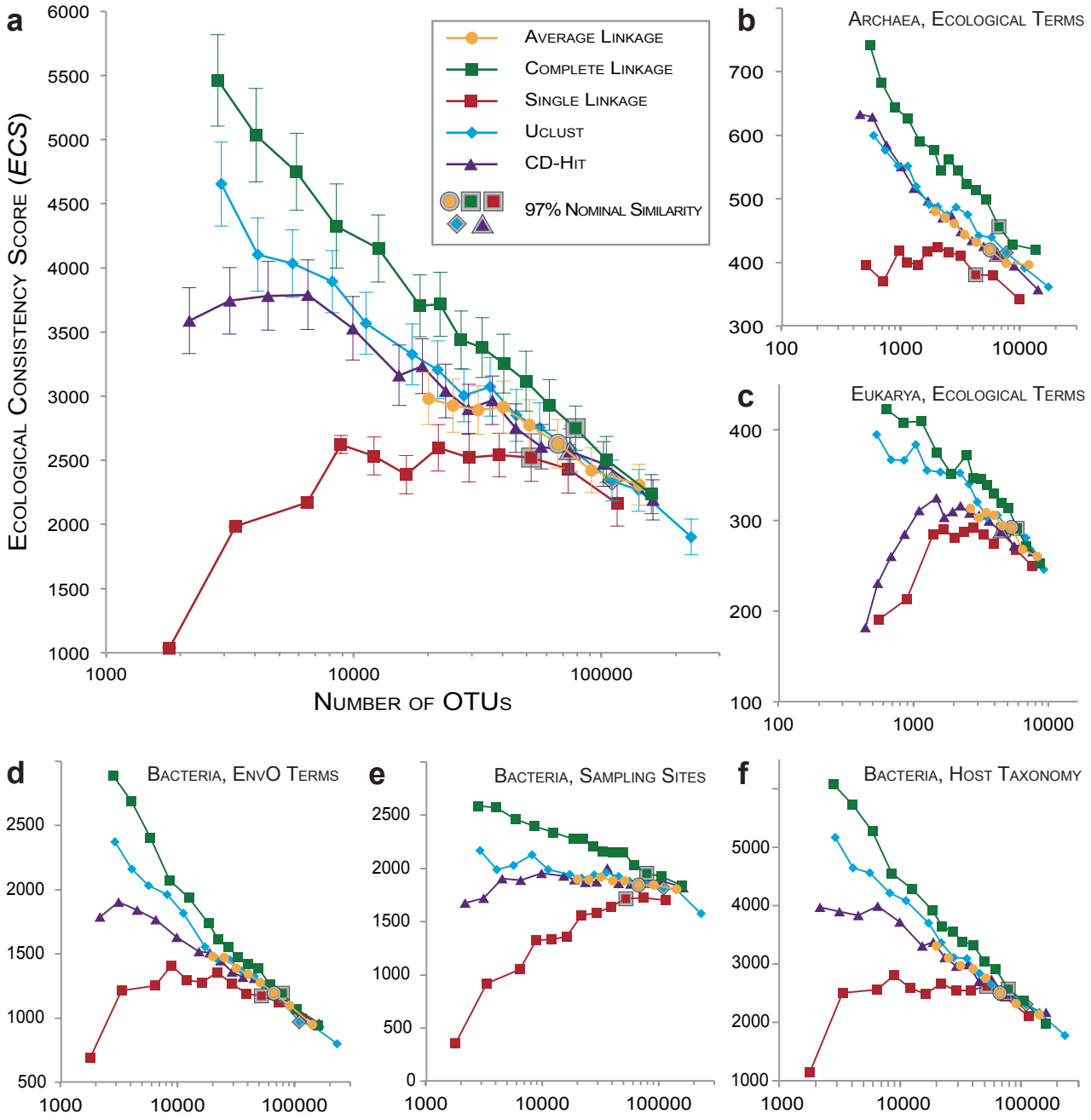


Figure 6.4. Global-scale ecological consistency of entire OTU sets. (a) Consistency of ecological terms across OTU sets when clustering 887,870 bacterial SSU sequences according to different methods. Error bars indicate jackknifed estimates of ECS variability. Data points for OTU sets clustered to 97% nominal sequence similarity are highlighted with a grey shade. For the ecological term consistency when clustering 42,402 archaeal sequences (b), or 20,120 eukaryotic 18S sequences (c), as well as for the bacterial dataset EnvO term consistency (d), sampling site consistency (e), and host taxonomy consistency (f), error bars are not drawn, but variability was in the same range as for (a) (coefficients of variation, 0.06 < c_V < 0.08). Figure reprinted with detailed discussion in manuscript 7.2.

Our results indicate that OTUs are generally, though not perfectly, ecologically consistent: at the granularities of the different tested ecological descriptions, enrichments of sequences sharing similar ecological affiliation were highly non-random. Considering that impartial OTU clustering is a largely theory-agnostic approach to defining diversity units, the observed levels of ecological consistency are remarkable: although OTUs make no assumptions on the evolution of 'true' bacterial taxa, they may still provide decent approximations of ecologically coherent lineages, at least at the given ecological resolution.

Koeppel & Wu recently reported "extensive ecological heterogeneity among OTUs" for very fine-scale habitat definitions of two model datasets [104]. Thus, OTU ecological consistency may in fact be a matter of perspective: while OTU demarcation may conflict with very high-resolution ecological associations for specific environments, we observed general ecological consistency at broader ecological scales.

Are these observed levels of ecological consistency relevant in practice? This question is additionally complicated by the recurrent observation that many recognized microbial taxa and many clusters of closely related organisms are *ecologically plastic*. While it has been shown that broad-scale ecological coherence in general is deeply rooted in phylogeny ([170]; an observation which we confirmed for our global SSU dataset, see Figure 1 in manuscript 7.2), several cases of wide 'intra-species' ecological variation have been reported, e.g. within the genera *Bacillus* [171] or *Escherichia* [118]. Thus, it is not trivial to define 'optimal' levels of ecological consistency in practice, in particular for the 'twilight zone' of uncharacterized, unnamed and uncharted microbial diversity. Nevertheless, our findings in general support the use of OTU-based approaches, as OTUs may provide 'good enough' approximations of ecologically coherent lineages to be useful in practice.

6.4 'How good is good enough?' – OTU ecological consistency and clustering 'quality'

The results presented in Figure 6.4 showed clear trends in ecological consistency between methods: *cl* generally provided highest, *sl* lowest levels of consistency, while *uclust* and *al/cd-hit* provided intermediate *ECS* values. These differences in ecological consistency may indeed be interpreted as differences in 'clustering quality', based on the following four main arguments. First, as discussed in the previous section, ecological coherence is one main criterion for 'true' microbial lineages. In consequence, one may infer that ecological consistency is a useful parameter to optimize for in OTU demarcation.

Second, there is a clear correlation between SSU sequence similarity and ecological similarity. This phenomenon has been described previously [170,172], and was confirmed on our global SSU dataset (see Figure 1a in manuscript 7.2). In general, the observation that closely related taxa ('species') share similar ecology, and in particular the evolutionary mechanisms underlying this phenomenon, are often referred to as *phylogenetic niche conservatism* [173]; in fact, this observation was already hinted at by Darwin [174]. At the same time, we could show for an example dataset that the reverse may also be true: ecologically more closely defined skin habitats in the HSM set shared higher levels of internal SSU similarity than expected for the global background dataset (Figure 1b in manuscript 7.2). In other words, while SSU similarity is generally indicative of ecological similarity ('more closely related organisms tend to be ecologically more similar'), the reciprocal statement may also be true: ecologically consistent groups of organisms may share higher levels of SSU sequence similarity than expected for a global background. For the problem of OTU clustering, this means that ecologically consistent OTUs may tend to cluster more closely related organisms, which indeed may be interpreted as 'good' clustering.

Third, microbial taxonomy reflects ecology: taxonomic labels often imply ecological (as well as morphological, or biochemical) descriptions of the groups of organisms they refer to. Although taxonomic labels may often fail to align with 'true' underlying diversity clusters, taxonomic typing of microbial ecology datasets remains important in practice – as taxonomic names confer an 'identity' to otherwise anonymous OTUs, and as taxonomic labels imply functional and ecological meaning. Thus, optimizing for OTU ecological consistency may likely facilitate coherent taxonomic descriptions of OTUs.

Lastly, ecologically consistent diversity units are desirable in microbial ecology applications: when investigating ecological properties of microbial communities, ecologically coherent OTUs will arguably be preferable to less consistent clusters.

Thus, the results shown in Figure 6.4 may be interpreted in terms of cluster 'quality'. In this view, *cl* clustering provided biologically more meaningful partitions than the other methods, while *sl* provided the most inconsistent OTU sets. As discussed in the previous section, differences in *ECS* are likely due to algorithm characteristics, rather than technical artifacts. Thus, the generally *exclusive* clustering regime applied by *cl* aligned well with sequence ecological affiliations, while *inclusive sl* clustering created clusters of conflicting ecology. Somewhat surprisingly, the average similarity-based *al* clusters achieved intermediate *ECS* levels. For the heuristic methods, the non-deterministic choice of cluster seed sequences may have generally reduced clustering accuracy, allowing ecologically heterogeneous sequence pairs to be clustered more frequently; seed effects have been described by Sun, Cai et al [58].

The observed trends may also be due to a differential influence of chimeric sequences on different clustering methods. Chimeric sequences are formed by template-switching during PCR amplification, and thus exhibit chimeric similarity patterns relative to their 'parent' strands [40]. Although we applied state-of-the-art chimera filtering which removed ~20% of all sequences from the dataset, the reliable identification and removal of chimeric sequences remains an open problem, so that a number of sequence chimeras may have remained in the dataset. In terms of sequence space, chimeric sequences share partial similarity with both their parent strands. As parent strands may share very low sequence similarity, chimeric sequences may thus lie 'in between' two otherwise unrelated sequence clusters, sharing intermediate similarity to both. Under *exclusive* clustering regimes, merging of two otherwise closely related clusters may therefore be prohibited if one of them contains a chimeric sequence that inflates furthest neighbor distance. In contrast, during *inclusive sl* clustering, chimeric sequences may serve as 'stepping-stones' that connect two otherwise very distant clusters by nearest-neighbor similarity – an effect which is arguably more detrimental. Such lumping of otherwise unrelated clusters may in part explain the low levels of ecological consistency observed for *sl*.

It is difficult to put these findings in the context of other benchmarks of OTU demarcation, not least because there is some discrepancy between previously reported trends; see section 3.4. For example, Schloss and Westcott [57] found that hierarchical *average linkage* clustering in *mothur* outperformed other methods in an internal benchmarking study. In contrast, Sun, Cai et al [71] found that their (slightly) heuristic *ESPRIT* implementation of *complete linkage* clustering outperformed *mothur*, later, however, Cai and Sun [72] and Sun, Cai et al [58] reported that their tool *ESPRIT-Tree*, implementing *al* clustering, outperformed all other tested methods in an *NMI*-based benchmark against taxonomic ground truth. These results were later corroborated by Bonder et al [66], who also recommended *uclust*, based on a benchmark for *NMI* and 'taxonomic purity'. In an earlier study, White et al [82] had found that hierarchical *cl* clustering, but in particular their proposed method '*Vl-cut*' outperformed other methods in a *Vl*-based test against taxonomic ground truth. In contrast, Huse et al [53] reported that *single linkage pre-clustering* to 99% similarity with subsequent *average linkage* clustering best reduced the 'overestimation' of OTU counts for a re-sequenced mock community; their results were later corroborated by Chen et al on simulated datasets [109].

Thus, while many studies have benchmarked OTU clustering, no single 'best' method stands out consistently. Nevertheless, several differences between previous tests and our approach are worth pointing out. First, as we have discussed above, differences in total OTU counts are not indicative of differences in cluster composition. Although most OTU count-based benchmark studies relied on well-defined ground truth datasets, differences

in total OTU count (or differential 'overestimation of diversity') may thus fail to indicate how well the ground truth partition was approximated. In other words, total cluster counts are arguably not a meaningful parameter to optimize for.

Second, several of the above-mentioned studies relied on taxonomic ground truths that were obtained using taxonomic classification tools such as the *RDP classifier* [59]. However, classification performance inherently depends on reference datasets, which are often biased towards well-studied taxa. Moreover, the use of taxonomic ground truth is arguably problematic in general due to the biased coverage of microbial diversity by existing taxonomic namespaces (see section 3.4.2).

Third, for any kind of ground truth-dependent benchmark, *NMI* and *VI* are probably unsuited measures: they provide shifting baseline values depending on differences in total cluster counts [120]. We note that none of the above studies corrected for this effect, but the general impact on interpretation is difficult to estimate.

Although the *ECS* approach presented here arguably circumvents some of these problems, it too suffers from shortcomings. For example, as sampling efforts have generally been biased towards selected environments (in particular, the human microbiome), the used dataset is likewise biased in ecological and phylogenetic coverage. Moreover, although different ecological signals and granularities were used, ecological resolution was not fully consistent across the entire sequence space, and did not reach the levels of detail required for very fine-scale descriptions. Nevertheless, we have presented what is to our knowledge the first benchmark of OTU demarcation that employs a signal *external* to both sequence and taxonomy. Moreover, ours is the first approach that covers a comprehensive, global survey of available microbial diversity data, so that our findings may be applicable beyond microbial ecology.

6.5 Conclusion

The work presented in this thesis revolved around the problem of OTU demarcation from complex sequencing datasets. Our findings elucidate the effects of OTU clustering on biological interpretation, and may enhance comparability and portability of results across studies. Thus, they may contribute to ongoing efforts towards more standardized sequence analysis pipelines. In general, in the debate on 'optimal' clustering methods, we adopted an outside perspective in manuscript 7.1: we investigated methods *relative* to each other, rather than benchmarking them against a potentially problematic ground truth. In manuscript 7.2, we have introduced an alternative approach towards assessing OTU quality – the optimization for ecological consistency, a biologically meaningful parameter.

In addition to the above discussion, there are a few conclusions to be drawn from the work presented. First, we note that our findings are relevant to reference-based OTU demarcation approaches, which have recently received increasing attention. In such protocols, sequences are mapped to curated reference OTU sets, to minimize the formation of spurious clusters. However, the accuracy of any such approach depends on the pre-clustering of the reference sequence set. The global dataset used in our study resembles in scope and level of pre-processing the frequently used reference datasets provided by the *RDP*, *Greengenes* and *SILVA* databases.

Second, several trends were consistent across all our tests. In particular, *uclust* (which is the default method in the widely used *QIIME* pipeline) was highly sensitive to clustering parameters and provided low reproducibility. In contrast, hierarchical *cl* clustering performed surprisingly well across tests: it was highly robust and reproducible, while also providing highest levels of ecological consistency.

Finally, our findings indicate that *de novo* OTU clustering is probably not as bad as its general reputation. Several high-profile research papers expressed a general skepticism towards OTUs, as they are arguably *proxies of proxies for proxies* (see above). However, we found that, depending on the choice of method, OTU clustering may robustly and reproducibly provide clusters that approximate ecologically coherent microbial lineages.

6.6 Outlook

In the so-called *technology hype cycle*, high-throughput sequencing-based approaches in microbial ecology are probably somewhere between the 'peak of inflated expectations' and the 'trough of disillusionment'. Similar to the great expectations related to the *human genome project* around the year 2000, there are at present great hopes associated with the recently completed *human microbiome project*, and to even more ambitious ongoing initiatives such as the *earth microbiome project*. When first released, however, the available human genome sequences revealed great gaps in our understanding that continue to require large efforts to be filled (such as e.g. the recently published *ENCODE* project). Similarly, microbiome research, and microbial ecology as such, have provided great insights, but they also revealed that the current understanding of the microbial world is narrowly confined. Moreover, the interpretation of contemporary high-throughput datasets on microbial communities poses several formidable problems. Current sequencing technology enables the study of almost any environment to almost arbitrary depths, yet not only computational analysis tools, but also scientific questions have arguably lagged behind. Many current studies that provide huge amounts of data are mostly descriptive – they report differences in microbial community composition between different habitats, or different treatments, etc. But a mere accumulation of more and deeper datasets does not enhance understanding, and principal component analysis too frequently remains the pinnacle of scientific endeavor when studying individual communities. Eventually, even the largest datasets will only be as informative as the research questions being asked. In order to unfold its true potential, I believe that microbial ecology will have to evolve from these data-driven approaches, back towards a real hypothesis-driven research field.

This transition is arguably happening already. Many recent studies use high-throughput sequencing to complement multiple lines of 'classical' microbial ecology evidence, rather than to replace them. At the same time, the continuing accumulation of data on a wide array of distinct environments also enables the study of global patterns and phenomena; e.g., there has recently been an increased interest in microbial co-occurrence networks generated from large meta-studies. Finally, increased data availability provides the context to identify and characterize the (few) remaining uncharted territories of microbial diversity.

Improved and more standardized computational analysis tools will be essential to these developments. In most approaches, OTU clustering plays an integral part, and will likely continue to do so. That is, until analyses at the level of individual sequences become computationally attainable – which would probably make sequence clustering superfluous, but which is unlikely to happen anytime soon given the current developments in sequencing technology. Similarly, high-throughput methodologies to study microbes at single-cell level are actively being developed, but their application to complex microbial ecology problems is still in its infancy. There are continuing efforts to further improve and refine clustering methods, e.g. by introducing additional (ecological) signals to inform sequence clustering. An objective assessment of methods, the development of impartial benchmarks and an optimization for biologically meaningful parameters will be essential to make informed choices, and to establish robust standards. Approaches as presented in this thesis have yet a part to play in this process.

environmental microbiology
environmental microbiology reports

Limits to Robustness and Reproducibility in the Demarcation of Operational Taxonomic Units

Journal:	*Environmental Microbiology and Environmental Microbiology Reports*
Manuscript ID:	Draft
Manuscript Type:	EMI - Research article
Journal:	Environmental Microbiology
Date Submitted by the Author:	n/a
Complete List of Authors:	Schmidt, Thomas; University of Zurich, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics Matias Rodrigues, João; University of Zurich, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics von Mering, Christian; University of Zurich, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics
Keywords:	bioinformatics, community genetics, environmental genomics, metagenomics/community genomics, microbial communities, microbial ecology, new tools/technological developments, uncultured microbes

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

Supplementary_Tables.tar.gz

Limits to Robustness and Reproducibility in the Demarcation of Operational Taxonomic Units

Thomas S. B. Schmidt[1], João F. Matias Rodrigues[1] & Christian von Mering[1,2]

[1] Institute for Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[2] to whom correspondence should be addressed (mering@imls.uzh.ch)

Summary

The demarcation of *Operational Taxonomic Units* (OTUs) from complex sequence datasets is a key step in contemporary studies of microbial ecology. However, as biologically motivated 'optimal' OTU binning algorithms remain elusive, many conceptually distinct approaches continue to be used. Using a global dataset of 887,870 bacterial 16S rRNA sequences, we objectively quantified biases introduced by several widely employed sequence clustering algorithms. We found that OTU-binning methods often provided surprisingly non-equivalent partitions of identical datasets, notably when clustering to the same nominal similarity thresholds; and we quantified the resulting impact on ecological data description for a well-defined human skin microbiome dataset. We observed that some methods were very robust to varying clustering thresholds, while others were found to be highly susceptible even to slight threshold variations. Moreover, we comprehensively quantified the impact of the choice of SSU gene subregion, as well as of dataset scope and context on algorithm performance. Our findings may contribute to an enhanced comparability of results across sequence processing pipelines, and we arrive at recommendations towards higher levels of standardization in established workflows.

Introduction

High-throughput sequencing technology has enabled the characterization of microbial communities at ever-increasing resolutions: individual environments have been probed to depths of millions of sequences, and even smaller-scale studies may routinely provide hundreds of thousands of reads. While cultivation-independent whole-genome sequencing has received increasing attention in the functional characterization of individual communities (The Human Microbiome Project Consortium, 2012a), targeted surveys for specific taxonomic marker genes, such as the 16S rRNA gene (Lane et al., 1985; Olsen et al., 1986), remain integral to many contemporary studies of microbial ecology. An essential first step in analyzing targeted 16S sequencing datasets is often the demarcation of basic units of diversity, ideally corresponding to 'true' microbial lineages that were present in the sample (Gevers et al., 2005; Cohan, 2006; Koeppel et al., 2008). However, in the absence of a unifying bacterial species concept (Doolittle and Papke, 2006; Achtman and Wagner, 2008; Doolittle and Zhaxybayeva, 2009), biologically motivated 'optimal' diversity unit definitions remain elusive, and a pragmatic approach is usually taken in practice: *Operational Taxonomic Units* (OTUs), defined as clusters of 16S sequence similarity, are used to approximate microbial taxa. Since OTU demarcation from complex 16S datasets is conceptually straightforward and often computationally efficient, OTUs are the backbone of established workflows for the ecological characterization of microbial communities, such as *mothur* (Schloss et al., 2009) and *QIIME* (Caporaso et al., 2010).

As the identification of 'optimal' partitions of large 16S datasets remains an open problem, a wide variety of OTU-binning methods have been developed. Traditionally, *hierarchical clustering algorithms* (implemented e.g. in *mothur*, *ESPRIT* (Sun et al., 2009) and *hpc-clust* (Matias Rodrigues and von Mering, 2014)) have been widely used, as have their *heuristic* approximations which include *cd-hit* (Li and Godzik, 2006; Fu et al., 2012), *uclust* (Edgar, 2010), *uparse* (Edgar, 2013), *ESPRIT-Tree* (Cai and Sun, 2011), *DySC* (Zheng et al., 2012), *MSClust* (Chen, Cheng, et al., 2013), *mBKM* (Wei et al., 2012) and *LSH* (Rasheed et al., 2013). While these methods rely on 'hard' clustering thresholds, several 'soft-threshold' or 'threshold-less' approaches have been proposed, including *CROP* (Hao et al., 2011), *M-Pick* (Wang et al., 2013) and *BEBaC* (Cheng et

al., 2012). Moreover, several algorithms rely on additional external data, either in the form of reference OTUs (e.g., 'reference-based OTU picking' strategies as implemented in *QIIME*), or in the form of additional biological signals, notably *ecotype simulation* (Koeppel et al., 2008), the tree-based *EPA-PTP* (Zhang et al., 2013) or *distribution-based clustering* (Preheim et al., 2013).

Given this diversity of available methods, several studies have aimed to identify those OTU-binning strategies which provide the 'best' partitions with respect to different objectives. The arguably most straightforward parameter to optimize for is the total number of clusters, as OTU counts serve as basis for estimates of community *richness*. Consequently, various studies have benchmarked OTU definitions based on total cluster counts, usually by assessing their overestimation of diversity with respect to test sets of known taxonomic composition, obtained by simulation or sequencing of mock communities (Sun et al., 2009; White et al., 2010; Huse et al., 2010; Schloss, 2010; Barriuso et al., 2011; Bonder et al., 2012; Chen, Zhang, et al., 2013). Other benchmarking strategies include dataset-internal quality measures (optimizing the ratio of *specificity* and *sensitivity* with respect to known input, e.g. by Schloss and Westcott, 2011; Li et al., 2012; Chen, Zhang, et al., 2013; Preheim et al., 2013) and external benchmarking against 'ground truth' datasets, optimizing for 'taxonomically pure' clusters (White et al., 2010; Sun et al., 2011; Cai and Sun, 2011; Bonder et al., 2012; Wang et al., 2013; Chen, Zhang, et al., 2013). More recently, OTU-binning methods have also been evaluated with respect to ecological consistency, by us and others (Koeppel and Wu, 2013; Schmidt et al., 2014).

Additional flexibility in 16S sequence processing workflows is introduced at several further levels, notably when removing sequencing noise and filtering for chimeric sequences (e.g., Schloss et al., 2011; Bonder et al., 2012), by sequence alignment strategies (White et al., 2010; Schloss, 2010; Barriuso et al., 2011; Sun et al., 2011; Wang et al., 2011; Schloss, 2012) and by sequence distance calculation (Schloss, 2010; Barriuso et al., 2011); however, these factors have been extensively discussed previously and are beyond the scope of our current study. Here, we are mainly concerned with the differences introduced by *algorithmic* choices, at the heart of the sequence clustering step.

Given the large flexibility at different levels, considerable efforts have been made to integrate and standardize workflows into one-stop-pipelines such as *mothur*, *QIIME* or *cd-hit-otu* (Li et al., 2012). However, in spite of these efforts and of a substantial body of literature on benchmarks, 'optimal' OTU demarcation strategies remain elusive, and the choice of methods and parameters varies considerably between studies. In consequence, it is generally difficult to compare ecological descriptions across studies, and study design is sometimes redundant, implementing complementary data analysis strategies to control for effects on biological interpretation – for example, the *human microbiome project* data was analyzed using multiple workflows, relying on *mothur* (*average linkage* clustering) and *QIIME-uclust* (The Human Microbiome Project Consortium, 2012b). Moreover, in spite of substantial efforts to benchmark OTU demarcation strategies, surprisingly little is known about the systematic differences *between* methods. It is not clear how similar approaches are in terms of resulting cluster *composition*, and how putative differences may bias biological interpretation beyond richness estimates. In other words, although OTU-binning strategies have been benchmarked extensively against varying concepts of 'optimality', systematic differences *between* methods are currently not well understood.

In this study, we explore limits to robustness and reproducibility in the demarcation of OTUs. We pursue a simple unifying question: how similar are different clustering methods? We approach this problem from various angles and quantify differences between five widely used clustering algorithms (*average, complete* and *single linkage* clustering, as well as the heuristics *cd-hit* and *uclust*) and for the recently published *uparse*, which implements adaptive on-the-fly chimera filtering. We selected these methods, because (i) they are capable of processing very large datasets, (ii) they are widely used in general and (iii) the rely on sequence data only (and not on external reference OTUs or additional phylogenetic or ecological signals to inform sequence clustering). We first revisit and quantify the observation that these methods generally provide diverging total cluster counts, and complement this earlier finding by investigating how these differences propagate to the level of cluster size distributions. We then turn our focus towards cluster *composition* and investigate how concordant the methods are when partitioning the very same set of sequences. In other words, do different methods tend to form consistent clusters, i.e. do they group similar sets of sequences? We first approach this question anecdotally, by re-

analyzing the well-studied *human skin microbiome* dataset for an individual clustering threshold, for which we explore how differences between methods translate to biases in ecological descriptions. We then broaden the scope of investigation to studying a global, comprehensive survey of publicly available full-length 16S sequence data across a wide range of clustering thresholds. In particular, we assess how robust methods are against slightly changing thresholds, and how reproducible partitions are across methods and thresholds. Finally, we assess robustness to changing clustering *context* (i.e., how does rich/sparse sequence space influence OTU demarcation?) and to the choice of 16S gene subregion.

Methods

Sequence data & preprocessing

We generated a comprehensive global SSU sequence dataset as described previously (Schmidt et al., 2014); see Text S1 for further details. In short, we parsed available full-length 16S rRNA gene sequences from NCBI GenBank (Benson et al., 2013) and from the genomes available in the NCBI Reference Sequence Database (RefSeq, Pruitt et al., 2011). After removing ~20% of total sequences that were flagged as chimeric by *uchime* (Edgar et al., 2011), we aligned the remaining sequences to a reference model for bacterial 16S (provided in the package *ssu-align*, Nawrocki, 2009) using *Infernal* (Nawrocki et al., 2009) and pruned away any terminal nucleotides that aligned outside of two manually chosen, well-conserved start- and end-positions. After these steps, our dataset comprised 887,870 aligned, near full-length bacterial 16S sequences.

From this global dataset, we extracted two smaller, 'local' datasets for in-depth analyses: the *human skin microbiome* (HSM) dataset (Grice et al., 2009), comprising 90,620 sequences after filtering and alignment; and an artificial dataset of *broad ecological range* (BER), combining 53,999 sequences from 18 studies focusing on distinct, unrelated environments (see Table S1).

Moreover, we generated three global datasets of *'short read'* sequences, by extracting subregions *V23* (pos 107-536 in the Infernal model, length 429nt, corresponding to 105-514 in the *E.coli* 16S sequence; *E.coli* reference positions as used by Schloss, 2010), *V35* (378-931, length 553nt, *E.coli* 357-906) and *V6* (1012-1072, length 60nt, *E.coli* 986-1045).

Sequence clustering into Operational Taxonomic Units

We clustered sequences into OTUs using three hierarchical clustering algorithms (*average, complete* and *single linkage*) and three heuristic methods (*cd-hit, uclust* and *uparse*). For every method, we clustered to varying thresholds between 90% and 100% sequence identity (in steps of 0.2%; 92-100% for *al*, 90-99% for *uclust* and *uparse*, see Text S1). We generated OTU sets using *cd-hit* (version 4.5.4, Build 2012-08-25, Fu et al., 2012) in *cd-hit-est* mode (default for the *cd-hit-otu* pipeline) from unaligned sequences using standard parameters. The *uclust* (version 6.0.307, Edgar, 2010) series of OTU sets was generated from unaligned sequences using the *uclust*

software with the *cluster_fast* option and standard parameters. As *uparse* (Edgar, 2013) combines OTU clustering with on-the-fly filtering for chimeric sequences, we used the full, unaligned, non-chimera-filtered sequence dataset for *uparse* runs and subsequently mapped shared sequences between *uparse* partitions and the *uchime*-filtered dataset used for the other clustering methods. Hierarchical *average, complete* and *single linkage* clustering were performed using our recently developed software package *hpc-clust* (Matias Rodrigues and von Mering, 2014), using the *onegap* sequence distance calculator (counting gaps as single mismatches). See Text S1 for additional details and parameters.

Assessing OTU set similarity

We assessed pairwise similarities between OTU sets using three distinct measures, namely *Normalized Mutual Information* (NMI, Fred and Jain, 2003), *Adjusted Mutual Information* (AMI, Vinh et al., 2009) and the *Adjusted Rand Index* (ARI, Hubert and Arabie, 1985); see Text S1 for further details. All three measures quantify the similarity in *cluster composition* between partitions (OTU sets): NMI, AMI and ARI values of 1 indicate perfectly identical clusterings.

Results

Quantitative differences between OTU definitions

When studying microbial communities, a crucial first step is often the characterization of local community complexity, richness and evenness, collectively referred to as *α-diversity*. Many measures of α-diversity rely on the total number of unique taxa observed in a sample (approximated by OTUs in practice), as well as their relative abundances. In consequence, several studies have used total cluster counts to benchmark OTU definitions (Sun et al., 2009; White et al., 2010; Huse et al., 2010; Schloss, 2010; Barriuso et al., 2011; Bonder et al., 2012; Chen, Zhang, et al., 2013).

To confirm and refine such previous observations, we clustered a global dataset of 887,870 near full-length bacterial 16S sequences using six different methods: *average linkage (al), complete linkage (cl), single linkage (sl), cd-hit, uclust* and *uparse*. We observed systematic shifts in total cluster counts between methods (Figure 1A; Table S2): when clustering to the same nominal sequence similarity threshold, *sl* provided the lowest, *uclust* the highest total OTU counts. All methods showed exponentially increasing counts with increasing clustering stringency, with over-exponential increases at very high similarities (≥98/99%). Interestingly, log-linear slopes were almost identical for *al*, *cl* and *cd-hit*, while *uparse* and *sl* diverged significantly; strikingly, the curve for *uclust* was not perfectly monotonous in the 96-97% threshold range.

These differences in overall cluster counts translated to differences in cluster size distributions. At a nominal similarity threshold of 97%, all tested methods provided differentially skewed OTU size histograms, with small OTUs (≤100 sequences) being notably overrepresented for *uclust* and underrepresented for *sl* (Figure 1B). Indeed, while *sl* clustered 78.8% of sequences into the largest 1.5% of OTUs (≥100 sequences), the largest 1% of *uclust* OTUs contained only 40.3% of total sequences; all other methods provided intermediate behavior (Figure 1C).

Differences in OTU composition between clustering methods

How do such differences in total cluster count and cluster size distribution translate to the level of individual OTUs? How similar are the various clustering methods with regard to the actual cluster composition? To approach these questions in a concrete example, we clustered 90,620 sequences of the well-studied *human skin microbiome* (HSM) dataset to 97% sequence similarity and traced sets of sequences (vertical bands) throughout the OTU sets (horizontal bars) in an alluvial flow diagram (Figure 2). Clearly, the tested OTU definitions provided markedly distinct partitions with respect to both cluster composition and cluster counts and sizes. While some clusters (e.g., *'OTU A'*) were almost identical between partitions, other sets of sequences showed characteristic behavior for the different clustering methods. *Single linkage* tended to produce large, comprehensive clusters (e.g., *'OTU B'*), lumping together sequences that would be split into multiple smaller OTUs by the other methods; this is in line with the generally *inclusive* single linkage algorithm (see also Figure 2, left panel). In contrast, both *cl* (e.g., *'OTU C'*) and *uclust* (e.g., *'OTU D'*) tended to split sequences into more and smaller clusters; in particular, these methods also clustered more sequences into 'small' OTUs (≤100 sequences, horizontal grey bars and dotted grey sequence band), which is in line with their 'splitting' behavior and the above observations on total cluster counts and size distributions. However, in spite of highly fluctuating partitions between methods, there were remarkably few cases of truly *discordant* clustering (sequences that completely traversed OTU boundaries), at least at the given resolution. Rather, differences between sets were almost always due to differential 'lumping' or 'splitting' of OTUs, although in some cases, the heuristic methods generated counterintuitive sub-partitions (e.g., sequences from *'OTU C'* clustered into one large *uclust* OTU that contained parts of several smaller *uparse* and *cd-hit* OTUs).

We used several measures to quantify the observed differences between sets (Figure 2, right panel, and Table 1): we assessed pairwise set similarity in terms of relative total OTU count (as binary log ratio, OC_{rel}), as well as in terms of cluster composition, using *Normalized Mutual Information* (NMI), *Adjusted Mutual Information* (AMI) and the *Adjusted Rand Index* (ARI; see Text S1). We observed that *uclust* and *sl* provided the most dissimilar partitions (NMI = 0.81, AMI =

0.80, ARI = 0.76, OC_{rel} = ±1.56), while *al* and *cd-hit* were the most similar (NMI = 0.96, AMI = 0.96, ARI = 0.95, OC_{rel} = ±0.11).

Qualitative differences between clustering methods may bias biological interpretation

To test how these differences in cluster counts and composition may influence biological interpretation, we re-analyzed the HSM dataset with respect to different ecological parameters. For the 21 skin sites sampled in the original HSM study, we estimated local diversity (α-diversity) based on three widely used measures: (i) the *Chao1* index, an abundance-based richness estimator which corrects for rare (unseen) classes (Chao, 1984); (ii) the *inverse Simpson* index, a classical abundance-weighted diversity measure (Simpson, 1949); and (iii) the entropy-based *Shannon* index (Shannon, 1948). Moreover, we assessed pairwise community similarity between habitats (β-diversity) using three different methods: (i) the abundance-informed *Sørensen-Dice-Czekanowski* (SDC) similarity index, which is closely related to the more well-known *Bray-Curtis* dissimilarity (Dice, 1945; Bray and Curtis, 1957); (ii) the *Morisita-Horn* (MH) overlap index (Horn, 1966); and (iii) Chao's *abundance-based Jaccard* (J_{abd}) index, which corrects for rare classes (Chao et al., 2004); see Text S1 for more details. The results across different clustering methods are shown in Figure 3 and Table S3.

We observed that clustering methods generally provided highly divergent α-diversity estimates: e.g., Chao1 richness estimates differed by up to 7.4-fold for individual samples ('inguinal crease', *uclust* versus *sl*, Figure 3A). Average shifts in diversity estimates were often systematic across samples and statistically significant (Figure 3B): *uclust* provided significantly higher diversity estimates than other methods (binary log ratio, 1.347-2.033 for Chao1, 0.073-0.853 for inverse Simpson, 0.292-0.642 for Shannon), while *sl* estimated systematically lower diversities and the other methods provided intermediate behavior. Nevertheless, all methods generally ranked the 21 samples similarly by diversity. In particular, *al, cl* and *cd-hit* provided very similar diversity trends (Pearson correlation across samples, 0.977-0.993 for Chao1, 0.969-0.989 for inverse Simpson and 0.991-0.998 for Shannon, Figure 3B), while for *uclust* and *uparse*, trends were considerably less

similar to other methods (Pearson correlation, lower limit of 0.434, Shannon index *uclust* vs *uparse*).

We observed similar effects for estimates of β-diversity. When comparing all pairwise community similarities between skin samples, *al*, *cl* and *cd-hit* provided very similar trends (Pearson correlation, 0.920-0.957 for SDC, 0.932-0.993 for MH, J_{abd} generally lower; Figure 3B), while *uclust*, *uparse* and *sl* provided lower correlations to other methods. Interestingly, MH estimates correlated very well (>0.9) between all methods except *uparse*, while J_{abd} provided comparatively low correlations (0.358-0.886); the latter effect is probably due to the correction for 'rare' (unseen) taxa implemented in J_{abd}, which will differentially distort community similarities according to OTU abundance distributions. SDC and J_{abd} estimates of community similarity were systematically lower for *uclust* (Mann-Whitney-U test, $p < 3*10^{-10}$) and higher for *sl* ($p < 1.4*10^{-14}$) when compared to all other methods. Systematic shifts in estimated community similarity between *al, cd-hit* and *cl* were far less pronounced, but sometimes statistically significant.

Thus, the choice of clustering method clearly had a significant impact on the ecological characterization of the HSM dataset; generally, *uclust*, *uparse* and *sl* deviated most in their descriptions of the data, while *al* and *cd-hit*, and to a lesser extent also *cl*, were more similar in diversity estimates between themselves. Next, we tested how these trends between clustering methods were captured by different measures of partition similarity (given in Table 2). In other words, we asked whether differences in cluster counts and cluster composition between methods may predict differences in ecological descriptions of the dataset. We found that in general, relative OTU counts between partitions (OC_{rel}) and ARI, AMI and NMI similarities correlated well with trends in Chao1 (Spearman correlation of pairwise partition similarities with diversity correlation across habitats, 0.861-0.886, Table 2). Correlations with trends in inverse Simpson, Shannon, SDC and J_{abd} were less pronounced, and trends in MH were moderately captured only by ARI similarity (Spearman correlation, 0.767). Systematic shifts in diversity estimates between methods in general corresponded well to trends in ARI, AMI and NMI for all diversity estimates except MH and inverse Simpson. Relative OTU counts were particularly poor indicators for shifts in the latter two estimators, but reasonably good indicators for the OTU count-correcting indices Chao1 and

J_{abd}, as expected. Thus, ARI, AMI and NMI similarities between partitions provided by different clustering methods were generally indicative both of differential trends in diversity estimates, as well as of their systematically shifted absolute values. In other words, differences between methods in cluster composition may in part explain biased diversity estimation between methods; when comparing two clusterings of the same data, ARI, AMI and NMI may in general indicate how biased these sets will be for downstream ecological descriptions.

General trends in robustness, reproducibility and similarity

In the previous sections, we have discussed how the choice of clustering method may influence biological interpretation of 16S sequence data. However, these findings are arguably anecdotal: they pertain to only one datapoint (clustering to 97% nominal similarity) for a defined model dataset. To generalize our observations, we clustered the 90,620 sequences in the HSM dataset to similarity thresholds between 90% and 100% (in steps of 0.2%) and calculated pairwise partition similarities for all combinations of clustering methods and thresholds as AMI, NMI and ARI (Figures S1-3; Tables S4-6). Moreover, to further broaden the scope of investigation, we performed a similar experiment on a global dataset of 887,870 bacterial 16S sequences, sampled from a wide array of environments (Figures 4, S4-5; Tables S7-9).

We observed that *al*, *cd-hit* and *cl* generally provided high partition similarities between themselves over wide cutoff ranges (AMI / NMI / ARI ≥ 0.9), while *sl*, *uclust* and *uparse* provided reasonably high partition similarities to other methods for the HSM dataset, but considerably lower similarities for the global dataset. In most cases, maximum partition similarities between two given methods were *off-diagonal*, indicating that nominal clustering thresholds were not directly equivalent across algorithms. For all methods, and at any threshold, best matches to *sl* partitions were shifted towards higher nominal thresholds for *sl*; in other words, when clustering e.g. to 97% similarity using *complete linkage*, the most similar *sl* partition was at >97% *sl* clustering. The reverse was true for *uclust*, and to a lesser extent *cl*: for these methods, maximum similarities to other methods were shifted towards lower nominal clustering thresholds (*e.g., al* clustering at 97% most similar to *cl* <97%). These effects are in line with the *inclusive* ('lumping') and *exclusive* ('splitting') nature of

the *sl* and *cl/uclust* algorithms. *Uparse* provided comparatively low similarities to all other methods across tested thresholds; this is likely due to *uparse*'s on-the-fly chimera filtering which removed different sets of sequences than our pipeline for the other methods – although we corrected for this effect by calculating partition similarity based only on shared sequences. Moreover, *uparse* reproducibly crashed when clustering to individual intermediate thresholds (such as 94%, 96% or 98.2%; see Text S1).

To test how robust clustering methods were against slight changes in similarity thresholds, we re-clustered the same datasets, but randomized the order of sequences (comparisons against 'self', diagonals in Figure 4). When clustering twice to the same threshold, all methods reproduced nearly identical partitions; note that for the heuristic methods, this is probably due to forced or internal sequence sorting, rather than deterministic algorithm behavior. However, even slight variations in threshold (increments of 0.2% corresponded to ~2.6 differences across 1301 alignment columns) had strong impacts on *uclust* and *uparse*. Effects on *cd-hit* and *sl* were less drastic, while *cl* and *al* were robust even to wider threshold variations.

We observed similar trends in robustness against varying thresholds when comparing partition similarities across methods. For *uclust* and *uparse*, similarity to other methods generally fluctuated with slightly changing clustering thresholds (vertically/horizontally 'striped' profiles in Figure 4), while similarities between the other tested methods were generally more robust. At very high similarity thresholds (≥99%), partition similarities dropped markedly for all methods, both when comparing between methods, as well as between different runs for the same method; this is likely due to very low levels of clustering at high stringencies (few sequences are clustered, most remain in singletons or very small OTUs).

As differences in total OTU counts have previously been used to benchmark clustering methods, we tested how predictive relative OTU counts were of differences in cluster composition (Spearman correlation of absolute binary log OTU count ratios, OC_{rel}, and AMI / NMI / ARI across thresholds, Table 3). We found that OC_{rel} correlated well with AMI, NMI and ARI for methods that provided generally similar partitions (e.g., *al* and *cd-hit*, correlation 0.858-0.945). However, for most pairwise comparisons of methods, only moderate or low correlations were observed: in

particular when comparing generally dissimilar methods, the difference in total OTU counts was a weak indicator of differences in cluster composition (e.g., *sl* and *uparse*, correlation 0.012-0.310). Indeed, the generally high AMI, ARI and NMI values across wide threshold ranges for some methods (in particular pairwise comparisons of *al*, *cl* and *cd-hit*) indicated that these methods provided partitions which were similar in spite of marked differences in OTU count. In other words, even though forming different total numbers of clusters, these methods tended to agree in OTU composition.

Finally, we assessed differential reproducibility between clustering methods, using partition similarities against other methods as a common reference. For all pairs of clustering methods, we correlated pairwise similarities to all other tested methods across thresholds; in other words, we asked how predictive the partition similarity of method A to a reference method X at a given threshold T was for the partition similarity of method B to method X at T% clustering (see also Text S1). We found that *al* and *cd-hit* behaved the most similarly (Spearman correlation, 0.991-0.993 for AMI and NMI; ARI generally lower; Table 4), while both methods were very similar to *cl* (0.853-0.915 and 0.830-0.899) and *uparse* (0.918-0.965 and 0.934-0.976). Moreover, *cl* also correlated well with *uclust* (0.892-0.924) and *uparse* (0.813-0.825). In contrast, *sl* provided low, and sometimes even slightly negative, correlations to other methods. We observed similar trends when assessing absolute differences in partition similarities relative to other methods (Figure S6). Notably, comparisons of relative OTU counts across methods provided very different correlations between methods, indicating that similar total OTU counts were generally not predictive of similar cluster composition.

A matter of perspective: sequence space and context affect OTU clustering

When comparing general trends in partition similarity for the well-defined HSM dataset (Figures S1-3) and the 'global' 16S set (Figures 4, S4-5), we observed that similarities between methods depended on the dataset: pairwise similarities and robustness to wide-range threshold changes were generally lower for the global set of 887,870 sequences than for the HSM. We hypothesized that these effects were due to differences in sequence space and context between the sets: the HSM set was arguably 'local', in the sense that it represented a more focused survey of microbial diversity than the comprehensive 'global' set.

To explore this hypothesis, and to quantify the differential impact of sequence context on clustering methods, we investigated two 'local' sets of sequences: (i) the HSM dataset, as described above; and (ii) an artificial dataset of 53,999 sequences from 18 samples of *broad ecological range* (BER; see Table S1). These sets covered very distinct sequence spaces: sequences in the HSM set shared significantly higher pairwise similarities than expected for the global set (Mann-Whitney-U test, $p << 10^{-16}$, Figure 5A), while for the BER set, sequences were significantly *less* similar ($p << 10^{-16}$). In other words, the HSM was indeed a more 'compact' subset of the global sequence set, while BER sequences were more dispersed, as illustrated in the toy sequence space representation in Figure 5.

We re-clustered both the HSM and BER sets twice – once in the presence, and once in the absence of the remaining sequences from the global set. We found that methods were differentially robust to clustering context, both in terms of total OTU counts (Figure 5B) and cluster composition (Figure 5C). While *al*, and to a lesser extent *cd-hit*, provided very similar cluster counts and compositions (AMI ≥ 0.9) across thresholds regardless of context, *sl* and in particular *uclust* were more strongly affected, providing up to 2-fold more (*sl*) or less (*uclust*) OTUs. *Complete linkage* provided diverging total cluster counts, but OTUs were highly similar by composition (AMI ≥ 0.94 at thresholds ≤ 98%). Generally, all methods except *sl* provided fewer OTUs under 'local' clustering, likely due to the absence of sequences which 'broke' OTUs into subclusters when partitioning a richer, global sequence space. In contrast, for *sl* a richer context would provide 'stepping-stone' sequences, connecting OTUs by closest-neighbor similarity which

remained separated under sparse, local context. Moreover, we observed that effects were generally more pronounced with decreasing thresholds (decreasing clustering stringency), except for *uclust*, which showed inverse behavior (higher AMI towards lower thresholds). Finally, *al*, *cd-hit* and *cl* showed a pronounced drop in partition similarity at very high thresholds (≥99%); note that *uclust* and *uparse* did not cluster the global set to these high stringencies.

We observed that effects were generally more pronounced for the HSM than for the BER set. Thus, the interpretation of focused datasets of ecologically similar samples, such as different skin habitats, may be more susceptible to clustering context than the arguably less realistic use-case of an ecologically broad set with dispersed sequence space. However, clustering context did have a significant impact in both scenarios.

Clustering methods are differentially robust to the choice of 16S gene subregion

Many contemporary studies in microbial ecology rely on sequencing of short, hypervariable subregions of the 16S gene, rather than of the full-length molecule, mostly for reasons of throughput and cost efficiency. To test how the choice of 16S subregion may affect clustering methods, we extracted datasets on subregions V23, V35 and V6 from the global alignment (see Methods and Figure 6A). While V23 and V35 approximated the subregions used in the *human microbiome project* (The Human Microbiome Project Consortium, 2012b), V6 has been a popular target in many Illumina-based studies (see e.g. Huse et al., 2010).

We found that partitions based on the different subregions generally diverged from clusterings of full-length 16S sequences. Across tested methods, V23 and V6 generally provided more OTUs than full-length 16S at lower thresholds, but fewer OTUs at higher stringencies, while V35 provided consistently fewer clusters (Figure 6B). Differences in OTU count were more pronounced for V6 than V23, while V35 was generally least affected; note that these results are in line with earlier findings by Schloss (2010) on different alignment and distance calculation methods for a smaller test dataset, and with findings by Kim et al (2011) on differences between subregions. The notable 'spikes' in V6 relative OTU counts and the generally discrete behavior for V23 and V35 were due

to sequence length effects at the given resolution: e.g. for the 60bp long V6, a single nucleotide mismatch corresponds to 1.67% sequence distance. Similarly, when clustering to the same nominal threshold, V23 and V35 were generally more similar to full-length 16S in cluster composition (AMI ≥ 0.8-0.9, Figure 6B) than V6, although partition similarity usually dropped markedly at very high thresholds (≥99%).

Clustering methods were differentially robust to the choice of 16S subregion. While *al* was least affected in terms of relative OTU counts, *al*, *cl*, *sl* and *uparse* were the most robust in terms of cluster composition. *Uclust* showed very similar (but comparatively low) partition similarities to full-length clustering for all tested subregions. When comparing partition similarities across varying thresholds (Figure 6C), we found that *al* and in particular *cl* were the most robust to changing stringencies, both on full-length and subregion clusterings; *sl*, and to a lesser extent *cd-hit* and *uparse*, were more susceptible, in particular for V6, while *uclust* provided generally lower similarities in cluster composition. Both *uparse* and *uclust*, and to a lesser extent *sl*, were susceptible to slight threshold changes in either subregion or full-length clusterings (horizontal/vertical 'stripes' in Figure 6C).

Thus, in spite of notable differences in total cluster counts, some tested methods were surprisingly robust to the choice of 16S subregion in terms of cluster composition, in particular *al* and *cl*. In other words, even when using shorter reads (containing less information), OTUs were often composed overall similarly, although there were clear differences between tested methods.

Discussion

Reproducibility of results is paramount to any empirical field of research. Scientific findings are generally required to be robust to the choice of experimental approach, and 'true' phenomena should be observable using independent methodologies, within reasonable limits. Drummond (2009) has formalized this notion by pointing out a conceptual distinction between the *replicability* of results and the *reproducibility* of findings. He contends that while the latter is a necessary prerequisite of scientific endeavor, the former is indeed less instructive. In other words, the exact replication of an experiment *ceteris paribus* is less informative than the corroboration of findings by reproduction in an independent setup (Casadevall and Fang, 2010).

We believe that these considerations are highly relevant to microbial ecology – which is not only an empirical, but indeed a data-driven research field. In this study, we have focused on the reproducibility of OTU demarcation from complex sequencing datasets: when repeating an experiment under different sequence clustering parameters, how much bias is introduced simply by the choice of methodology? In other words, how robust are biological findings to the choice of clustering method? We found that OTU demarcation may indeed be *replicable*: different methods provided (almost) identical partitions when twice clustering the exact same sets of sequences, but in randomized order (Figure 4, diagonals). However, trends in *reproducibility* were less clear.

We quantified the variability in OTU demarcation on various complementary levels. In a first, basic approach, we confirmed previous observations on diverging total OTU counts and cluster size distributions between methods (Figure 1). However, total cluster counts are a summary statistic of limited biological significance with respect to OTU composition (Table 3) and higher-level ecological data descriptions (Figure 3, Table 2). Rather, differences between clusterings are more meaningfully described as differences in cluster *composition*. We explored these on an anecdotal level for an exemplary datapoint (Figure 2), for which we also quantified biases to higher level ecological data descriptions (Figure 3). We then generalized our observations to a global dataset, and to the choice of clustering threshold (Figures 4, S1-5), clustering context (Figure 5) and 16S sequence subregion (Figure 6).

When viewed across all these tests, hierarchical *average linkage* and heuristic *cd-hit* clustering were generally the most similar pair of methods. This is both surprising and remarkable, as *cd-hit* relies on several computationally efficient shortcuts which are expected to reduce accuracy. Both *al* and *cd-hit* also showed generally similar behavior to *complete linkage*. Similarities in cluster composition between these three methods were robust to (wide) changes in clustering threshold, indicating that these methods provided surprisingly reproducible clusterings. These high levels of similarity between *al*, *cl* and *cd-hit* are remarkable, in particular when considering that these methods diverged considerably in terms of total OTU counts across thresholds. In contrast, *single linkage*, *uclust* and *uparse* diverged more strongly in their behavior from all other methods. Indeed, the 'inclusive' *single linkage* algorithm is a conceptual outlier in the tested set of methods, as it implements a fundamentally different clustering regime than the other, more 'exclusive' methods. Similarly, *uparse* can be considered an outlier, as it implements adaptive on-the-fly chimera filtering and effectively clustered different sets of sequences than the other tested methods. Indeed, *uparse* filtering for chimeric sequences was far more restrictive than for the *uchime*-based protocol (with *uparse* removing ≥50% sequences for some thresholds), which is surprising when considering that the input data were full length, high-quality and often curated or pre-filtered 16S sequences. Thus, lower similarities of *uparse* relative to other methods in cluster composition, as well as in overall behavior could be expected. In contrast, though conceptually related to *cd-hit* and *cl*, *uclust* clearly diverged across many tests: with respect to other methods, it provided significantly shifted diversity estimates (Figure 3) and deviating cluster composition, in particular at higher (biologically more relevant) clustering thresholds (Figure 4).

Although it is tempting to interpret these findings in terms of *cluster quality*, we note that notion of 'true' or 'false' clustering requires more than pairwise comparisons between methods, and has been addressed elsewhere, by us and others (e.g., Sun et al., 2011; Koeppel and Wu, 2013; Schmidt et al., 2014). Rather, partition similarities across thresholds may inform the comparison of results across studies, as they allow an assessment of the bias introduced by clustering method. In contrast, the observed trends in *robustness* to changing parameters may indeed be interpreted in either way: in terms of comparability across studies, but also as (quality) attributes of clustering methods. In particular, *cl*, *al* and *cd-hit* were surprisingly robust to changes in clustering threshold,

clustering context and the choice of subregion, while *sl*, *uparse* and especially *uclust* were more strongly affected. By design, previous unidimensional benchmarking studies focusing on different concepts of partition 'optimality' did not capture these trends in robustness between methods.

There have been great efforts to address the 'reproducibility' issue through increased levels of standardization: software pipelines such as *mothur* or *QIIME* provide comprehensive protocols to analyze microbial ecology datasets. However, these efforts have arguably enhanced *replicability* rather than *reproducibility*, by providing widely adopted defaults. Furthermore, we note that *QIIME* relies on *uclust* as default clustering method – the method which was consistently the most sensitive to any parameter change across our tests. Thus, while *QIIME* aims to enhance comparability of findings across studies, at the level of sequence clustering it probably achieves the reverse. In contrast, the default clustering method implemented in *mothur* is *al* which was among the most robust methods in our tests.

Reference-based OTU demarcation is another approach to standardization which has recently received increasing attention: sequences are mapped to pre-clustered reference sets of curated 16S sequences, provided e.g. by the RDP (Cole et al., 2013), Greengenes (DeSantis et al., 2006) and SILVA (Yilmaz et al., 2013) databases. We note that the global dataset used in our study closely resembles such reference sets in size, scope, sequence length and pre-processing. Moreover, one main difference between reference-based OTU demarcation and *de novo* clustering is arguably clustering context – an effect which has previously been ignored or underestimated. Most importantly, however, the 'quality' of reference-picked OTUs directly depends on the quality of pre-clustering of the reference set. The Greengenes and SILVA databases, which are the default reference sets in *QIIME* and for the *earth microbiome project* (Gilbert et al., 2010), are pre-clustered to 97% and 99% similarity using *uclust*.

In view of the many parameter choices in sequence processing pipelines, how can reproducibility of results be enhanced in practice? Based on our findings, we suggest that researchers may want to resort to deliberately redundant study/analysis designs. Several recent studies, e.g. the *human microbiome project*, have indeed relied on complementary analysis pipelines, but this is usually not the case. Redundancy may be introduced at many levels, e.g. in the choice of sequenced

subregions, and at every level of sequence processing, and we recommend that researchers implement at least two complementary analysis pipelines. Biological findings that are robust to independent methodologies are arguably more dependable than any single-track analysis.

Acknowledgements

We thank Mark Robinson, University of Zürich, for insightful discussions, and for pointing us to the Adjusted Rand Index, as well as Damian Szklarczyk and Alexander Roth for helpful discussions during the preparation of the manuscript. This work was supported by an ERC starting grant to CvM (UMICIS/242870), and by the Swiss National Science Foundation (31003A_135688). The authors declare that no conflict of interest exists.

References

Achtman, M. and Wagner, M. (2008) Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* **6**: 431–440.

Barriuso, J., Valverde, J.R., and Mellado, R.P. (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics* **12**: 473.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013) GenBank. *Nucleic Acids Research* **41**: D36–42.

Bonder, M.J., Abeln, S., Zaura, E., and Brandt, B.W. (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics (Oxford, England)* **28**: 2891–2897.

Bray, J.R. and Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27**: 325–349.

Cai, Y. and Sun, Y. (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research* **39**: e95.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–336.

Casadevall, A. and Fang, F.C. (2010) Reproducible Science. *Infection and Immunity* **78**: 4972–4975.

Chao, A. (1984) Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* **11**: 265–270.

Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.-J. (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8**: 148–159.

Chen, W., Cheng, Y., Zhang, C., Zhang, S., and Zhao, H. (2013) MSClust: A Multi-Seeds based Clustering algorithm for microbiome profiling using 16S rRNA sequence. *Journal of Microbiological Methods* **94**: 347–355.

Chen, W., Zhang, C.K., Cheng, Y., Zhang, S., and Zhao, H. (2013) A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLOS ONE* **8**: e70837.

Cheng, L., Walker, A.W., and Corander, J. (2012) Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research* **40**: 5240–5249.

Cohan, F.M. (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos T R Soc B* **361**: 1985–1996.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., et al. (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**: D633–D642.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dice, L.R. (1945) Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**: 297–302.

Doolittle, W.F. and Papke, R.T. (2006) Genomics and the bacterial species problem. *Genome Biol* **7**: 116.

Doolittle, W.F. and Zhaxybayeva, O. (2009) On the origin of prokaryotic species. *Genome Res* **19**: 744–756.

Drummond, C. (2009) Replicability is not reproducibility: nor is it good science. Montreal, Quebec, Canada.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)* **26**: 2460–2461.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Publishing Group* **10**: 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)* **27**: 2194–2200.

Fred, A.L.N. and Jain, A.K. (2003) Robust Data Clustering. pp. 128–136.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**: 3150–3152.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., et al. (2005) Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**: 733–739.

Gilbert, J., Meyer, F., Antonopoulos, D.A., Balaji, P., Brown, C.T., Brown, C.T., et al. (2010) Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* **3**: 243–248.

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., et al. (2009) Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324**: 1190–1192.

Hao, X., Jiang, R., and Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics (Oxford, England)* **27**: 611–618.

Horn, H.S. (1966) Measurement of“ overlap” in comparative ecological studies. *American Naturalist* 419–424.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification* **2**: 193–218.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**: 1889–1898.

Kim, M., Morrison, M., and Yu, Z. (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods* **84**: 81–87.

Koeppel, A., Perry, E.B., Sikorski, J., Krizanc, D., Warner, A., Ward, D.M., et al. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci USA* **105**: 2504–2509.

Koeppel, A.F. and Wu, M. (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research* **41**: 5175–5188.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**: 6955–6959.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)* **22**: 1658–1659.

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics* **13**: 656–668.

Matias Rodrigues, J.F. and von Mering, von, C. (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics (Oxford, England)* **30**: 287–288.

Nawrocki, E.P. (2009) Structural RNA Homology Search and Alignment Using Covariance Models. 1–281.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)* **25**: 1335–1337.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**: 337–365.

Preheim, S.P., Perrotta, A.R., Martin-Platero, A.M., Gupta, A., and Alm, E.J. (2013) Distribution-Based Clustering: Using Ecology to Refine the Operational Taxonomic Unit. *Appl Environ Microbiol* **79**: 6593–6603.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research* **40**: D130–D135.

Rasheed, Z., Rangwala, H., and Barbará, D. (2013) 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. *BMC systems biology* **7**: S11.

Schloss, P.D. (2012) Secondary structure improves OTU assignments of 16S rRNA gene sequences. *ISME J* **7**: 457–460.

Schloss, P.D. (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLOS Computational biology* **6**: e1000844.

Schloss, P.D. and Westcott, S.L. (2011) Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl Environ Microbiol* **77**: 3219–3226.

Schloss, P.D., Gevers, D., and Westcott, S.L. (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLOS ONE* **6**: e27310.

Schloss, P.D., Westcott, S.L., Rabyn, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75**: 7537.

Schmidt, T.S.B., Matias Rodrigues, J.F., and von Mering, von, C. (2014) Ecological Consistency of SSU rRNA-based Operational Taxonomic Units at a Global Scale. *PLOS Computational biology* 10: e1003594

Shannon, C.E. (1948) A Mathematical Theory of Communication. *At&T Tech J* **27**: 623–656.

Simpson, E.H. (1949) Measurement of Diversity. *Nature* **163**: 688–688.

Sun, Y., Cai, Y., Huse, S.M., Knight, R., Farmerie, W.G., Wang, X., and Mai, V. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics* **13**: 107–121.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W., and Farmerie, W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research* **37**: e76–e76.

The Human Microbiome Project Consortium (2012a) A framework for human microbiome research. *Nature* **486**: 215–221.

The Human Microbiome Project Consortium (2012b) Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207–214.

Vinh, N.X., Epps, J., and Bailey, J. (2009) Information theoretic measures for clusterings comparison. ACM Press, New York, NY, pp. 1073–1080.

Wang, X., Cai, Y., Sun, Y., Knight, R., and Mai, V. (2011) Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences. *ISME J* **6**: 1277–1280.

Wang, X., Yao, J., Sun, Y., and Mai, V. (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics* **14**: 43.

Wei, D., Jiang, Q., Wei, Y., and Wang, S. (2012) A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics* **{13}**:.

White, J.R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M. (2010) Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics* **11**: 152.

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2013) The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research* **42**: D643–D648.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics (Oxford, England)* **29**: 2869–2876.

Zheng, Z., Kramer, S., and Schmidt, B. (2012) DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics (Oxford, England)* **28**: 2182–2183.

Table & Figure Legends

Figure 1. Quantitative differences between clustering methods. (A) Differences in total OTU counts when clustering a global dataset of 887,870 bacterial 16S sequences according to different methods. Note that *uparse* filtered for chimeric sequences differently than the other methods, which led to different numbers of sequences being clustered at different cutoffs (see Table S2). Moreover, *uclust* and *uparse* did not cluster to >99% similarity, with additional missing datapoints for *uparse* (see Text S1). (B) Differences in OTU size distributions between methods when clustering to 97% nominal sequence similarity. (C) Differential dominance of singleton and large OTUs (≥100 sequences) at 97% similarity. Methods differed in the fraction of total sequences (upper panel) and of total OTUs (lower panel) that fell into different OTU size categories.

Figure 2. Differences in OTU composition at an individual datapoint. 90,620 bacterial 16S sequences of the *human skin microbiome* (HSM) dataset were clustered to 97% sequence similarity according to different methods (algorithms illustrated in left panel); note that here, we additionally used the '-id 0.97' option for *uparse*. Sets of sequences (vertical bands) were clustered into OTU sets (horizontal bars) differentially between methods. Coloring highlights separate sequence bands; the dotted grey band corresponds to small OTUs (≤100 sequences). Partition similarities (right panel) were quantified in terms of relative OTU counts (as binary logarithm), AMI, NMI and ARI values; see also Table 1.

Figure 3. Clustering methods introduce differential bias to ecological data descriptions. (A) Chao1 richness estimates for the 21 samples of the HSM dataset when clustering to 97% sequence similarity using different methods. Vertical grey bars indicate median values. (B) Biases to estimators of local diversity (*Chao1, inverse Simpson* and *Shannon indices*) and community similarity (*Sørensen-Dice-Czekanowski, Morisita-Horn* and *abundance-corrected Jaccard indices*; see Text S1) for the same 21 skin samples. For every pair of clustering methods, Pearson correlations (upper cells) indicate differences in diversity ranking between samples per index. Average shifts in absolute diversity estimates per index ('relative shift', lower cells) between

methods are calculated 'row-wise' (e.g., for the highlighted cell, read '*sl* provided lower Chao1 estimates than *cl*'). Cell coloring indicates significance levels and direction of absolute shifts between methods (Mann-Whitney-U test). Raw data on diversity estimates is provided in Table S3.

Figure 4. Differences in cluster composition between methods across wide threshold ranges. A global dataset of 887,870 16S sequences was clustered to thresholds ranging from 90-100% sequence similarity, in steps of 0.2% (corresponding to ~2.6 differences across the full sequence length). Pairwise partition similarities between methods and across thresholds were calculated as *Adjusted Mutual Information* (AMI). To calculate partition similarities of methods 'against themselves' (subplots on diagonal), clustering was re-run with randomized order of sequences. Note that algorithm memory requirements prohibited *al* clustering of the full set to <92% similarity and *uclust/uparse* clustering to >99% similarity; moreover *uparse* consistently crashed when clustering the dataset to 96.0% or 98.2% similarity (grey lines in the corresponding plots). Equivalent plots on NMI and ARI similarities, and for the HSM dataset, are provided as Figures S1-5. Raw data on partition similarities provided in Tables S4-9.

Figure 5. Robustness to clustering context. (A) The HSM and an artificially generated dataset of *broad ecological range* (BER, see Table S1) were extracted as 'local' subsets from the global set of 887,870 16S sequences. Pairwise internal sequence similarities were calculated based on 10 randomly drawn sets of 10,000 sequences per dataset. Compared to the global background, internal sequence similarities were significantly higher for the HSM set, and significantly lower for the BER set ($p << 10^{-16}$, Mann-Whitney-U test). This corresponds to the patterns of filled dots (HSM/BER) over circles (global background) in the lower panel in (A). (B) Relative OTU counts when clustering HSM and BER sets in the presence ('global context') and absence ('local context') of the full global sequence space. (C) Partition similarities between local and global context, expressed as *Adjusted Mutual Information* (AMI).

Figure 6. Robustness to the choice of SSU gene subregion. (A) Extraction of selected hypervariable subregions from near full-length 16S sequence alignments ('FL'). Sequence length (left column) and nucleotide positions in the *ssu-align* bacterial 16S model (middle column, Nawrocki, 2009), and the *E. coli* reference 16S sequence (right column) for subregions *V23, V35* and *V6* were chosen following Schloss (2010), and indicated in the secondary structure resolved *E. coli* 16S sequence on the right (modified from an image kindly provided by Harry Noller, University of California, Santa Cruz). (B) Relative OTU counts of subregion clustering over full-length clustering and partition similarities as *Adjusted Mutual Information* (AMI) when clustering to the same nominal similarity threshold according to different methods. (C) Partition similarities across clustering thresholds reveal differential trends in robustness between methods.

Table 1. Pairwise partition similarities at an individual datapoint. The HSM dataset was clustered to 97% nominal sequence similarity according to different methods (see also Figure 2), and pairwise partition similarities were calculated as relative OTU counts (row-wise, as binary log-ratios, i.e. read '*cl* provided $2^{0.261}$ times as many OTUs as *al*'), AMI, NMI and ARI. Full pairwise partition similarities across thresholds are available in Tables S4-6.

Table 2. Partition similarities may predict trends in ecological data description. Pairwise partition similarities between methods (OC_{rel}, AMI, NMI and ARI), as shown in Table 1, were correlated with trends and shifts in α- and β-diversity estimates between methods across skin habitats (shown as Pearson correlations and relative shifts in Figure 3B). In other words, each value in the table indicates how well partition similarities correlated with similarities in ecological data descriptions (Spearman rank correlation; see Text S1 for further details).

Table 3. Differences in OTU counts are a poor predictor of differences in cluster composition. For every pair of methods, differences in OTU counts (as absolute binary log ratios) and in cluster composition (as AMI, NMI and ARI) were correlated across thresholds (Spearman rank correlation). In other words, every value in the table indicates how predictive differences in

OTU counts were of (AMI / NMI / ARI) differences in cluster composition. See Text S1 for further details.

Table 4. Pairwise similarities between clustering methods, expressed as shared trends in partition similarities to other methods. For every pair of methods, partition similarities (as OC_{rel}, AMI, NMI and ARI) across all methods and thresholds were correlated (Pearson correlation). For example, the value for AMI correlations between *al* and *cl* was calculated from pairwise AMI similarities of *al* to all methods across thresholds, which were correlated to AMI similarities of *cl* partitions across methods and thresholds. In other words, values in the table indicate how similarly two methods behave, using partition similarities across methods and thresholds as common reference.

Supporting Information Legends

Text S1. Supplementary Methods.

Figure S1. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the HSM dataset. Equivalent to Figure 4 in the main text. Raw AMI values provided in Table S4.

Figure S2. Normalized Mutual Information (NMI) between methods across thresholds when clustering the HSM dataset. Equivalent to Figure 4 in the main text. Raw NMI values provided in Table S5.

Figure S3. Adjusted Rand Index (ARI) between methods across thresholds when clustering the HSM dataset. Equivalent to Figure 4 in the main text. Raw ARI values provided in Table S6.

Figure S4. Normalized Mutual Information (NMI) between methods across thresholds when clustering the global dataset of 887,870 16S sequences. Equivalent to Figure 4 in the main text. Raw NMI values provided in Table S8.

Figure S5. Adjusted Rand Index (ARI) between methods across thresholds when clustering the global dataset of 887,870 16S sequences. Equivalent to Figure 4 in the main text. Raw ARI values provided in Table S9.

Figure S6. Pairwise similarities between clustering methods, expressed as absolute differences in partition similarities to other methods. For every pair of clustering methods, differences in partition similarities (expressed as AMI) to all methods across thresholds are shown as histograms. For example, the top left subgraph shows differences between *al* and all other methods; it shows that *cd-hit* and *al* provide very similar AMI values against other methods across

thresholds, although *cd-hit* AMI values tend to be slightly lower (peak shifted to the left). In other words, the subplots indicate how similarly pairs of methods behave, using partition similarities to other methods across thresholds as reference.

Table S1. Composition of an artificial 'local' dataset of *broad ecological range* (BER). A total of 53,999 16S sequences from 18 studies were selected to generate the dataset; additional information, and detailed references are provided in the table.

Table S2. Total OTU counts per method when clustering a global dataset of 887,870 16S sequences. OTU counts are given per method for different thresholds. For *uparse*, the respective number of clustered sequences that mapped to the *uchime*-filtered dataset is also provided (as *uparse* implements differential on-the-fly chimera filtering, removing different sets of sequences at different clustering thresholds).

Table S3. Estimates of α- and β-diversity when clustering the HSM dataset to 97% sequence similarity according to different methods. The table provides raw data of diversity estimates across 21 skin samples for every method. Moreover, trends between methods (Pearson correlation of diversity estimates), absolute shifts (as binary log-ratio) and shift significance (as p-values in one-sided Mann-Whitney-U tests) are also provided for every index. Finally, an overview of the HSM dataset (sequence counts and internal sequence similarities per habitat), as well as pairwise partition similarities between methods at 97% clustering are also provided.

Table S4. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the HSM dataset. Values as shown in Figure S1.

Table S5. Normalized Mutual Information (NMI) between methods across thresholds when clustering the HSM dataset. Values as shown in Figure S2.

Table S6. Adjusted Rand Index (ARI) between methods across thresholds when clustering the HSM dataset. Values as shown in Figure S3.

Table S7. Adjusted Mutual Information (AMI) between methods across thresholds when clustering the global dataset of 887,870 16S sequences. Values as shown in Figure 4 in the main text. Missing values for *al* at thresholds <92% and *uclust / uparse* >99% as clustering to these thresholds was prohibited by memory requirements (see Text S1).

Table S8. Normalized Mutual Information (NMI) between methods across thresholds when clustering the global dataset of 887,870 16S sequences. Values as shown in Figure S4. Missing values for *al* at thresholds <92% and *uclust / uparse* >99% as clustering to these thresholds was prohibited by memory requirements (see Text S1).

Table S9. Adjusted Rand Index (ARI) between methods across thresholds when clustering the global dataset of 887,870 16S sequences. Values as shown in Figure S5. Missing values for *al* at thresholds <92% and *uclust / uparse* >99% as clustering to these thresholds was prohibited by memory requirements (see Text S1).

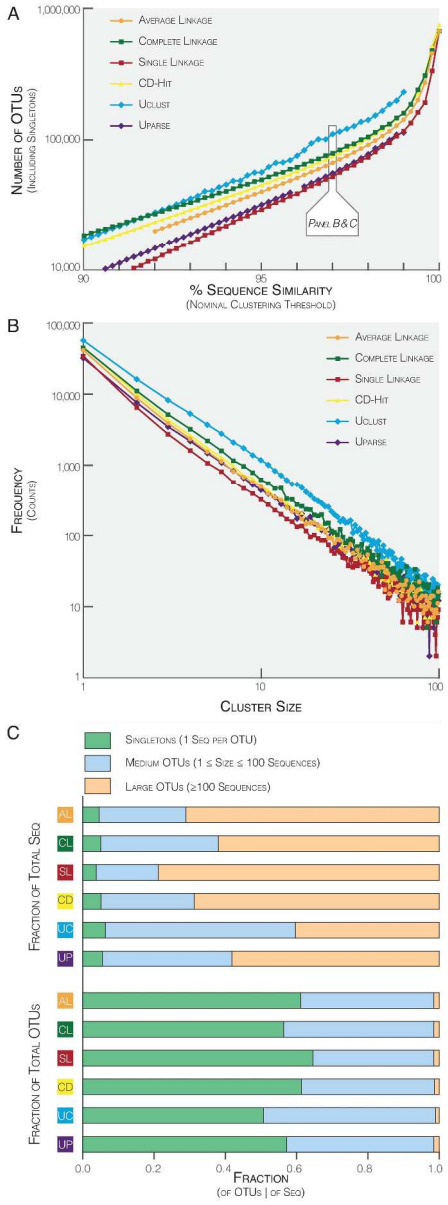


Figure 1. Quantitative differences between clustering methods. (A) Differences in total OTU counts when clustering a global dataset of 887,870 bacterial 16S sequences according to different methods. Note that uparse filtered for chimeric sequences differently than the other methods, which led to different numbers of sequences being clustered at different cutoffs (see Table S2). Moreover, uclust and uparse did not cluster to >99% similarity, with additional missing datapoints for uparse (see Text S1). (B) Differences in OTU size distributions between methods when clustering to 97% nominal sequence similarity. (C) Differential dominance of singleton and large OTUs (≥100 sequences) at 97% similarity. Methods differed in the fraction of total sequences (upper panel) and of total OTUs (lower panel) that fell into different OTU size categories.
229x629mm (300 x 300 DPI)

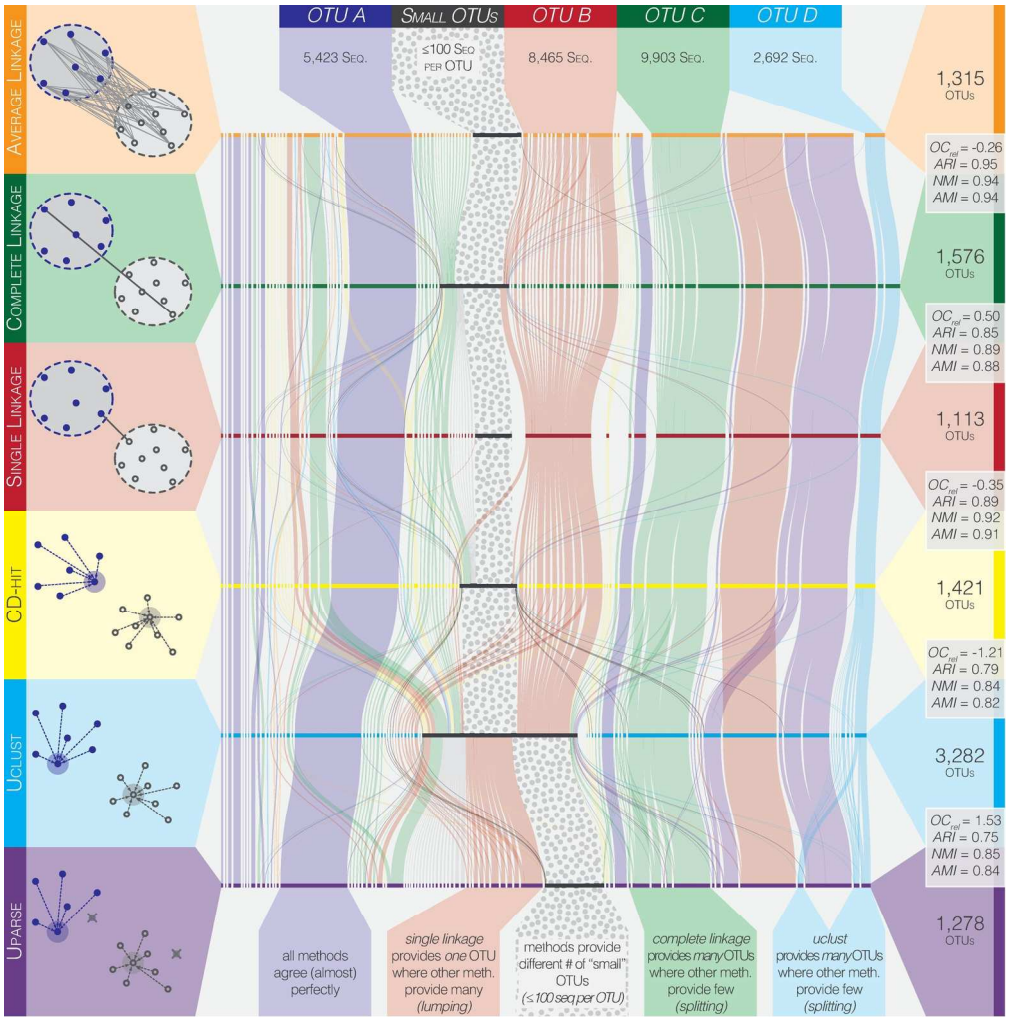


Figure 2. Differences in OTU composition at an individual datapoint. 90,620 bacterial 16S sequences of the human skin microbiome (HSM) dataset were clustered to 97% sequence similarity according to different methods (algorithms illustrated in left panel); note that here, we additionally used the '-id 0.97' option for uparse. Sets of sequences (vertical bands) were clustered into OTU sets (horizontal bars) differentially between methods. Coloring highlights separate sequence bands; the dotted grey band corresponds to small OTUs (≤100 sequences). Partition similarities (right panel) were quantified in terms of relative OTU counts (as binary logarithm), AMI, NMI and ARI values; see also Table 1.
180x181mm (300 x 300 DPI)

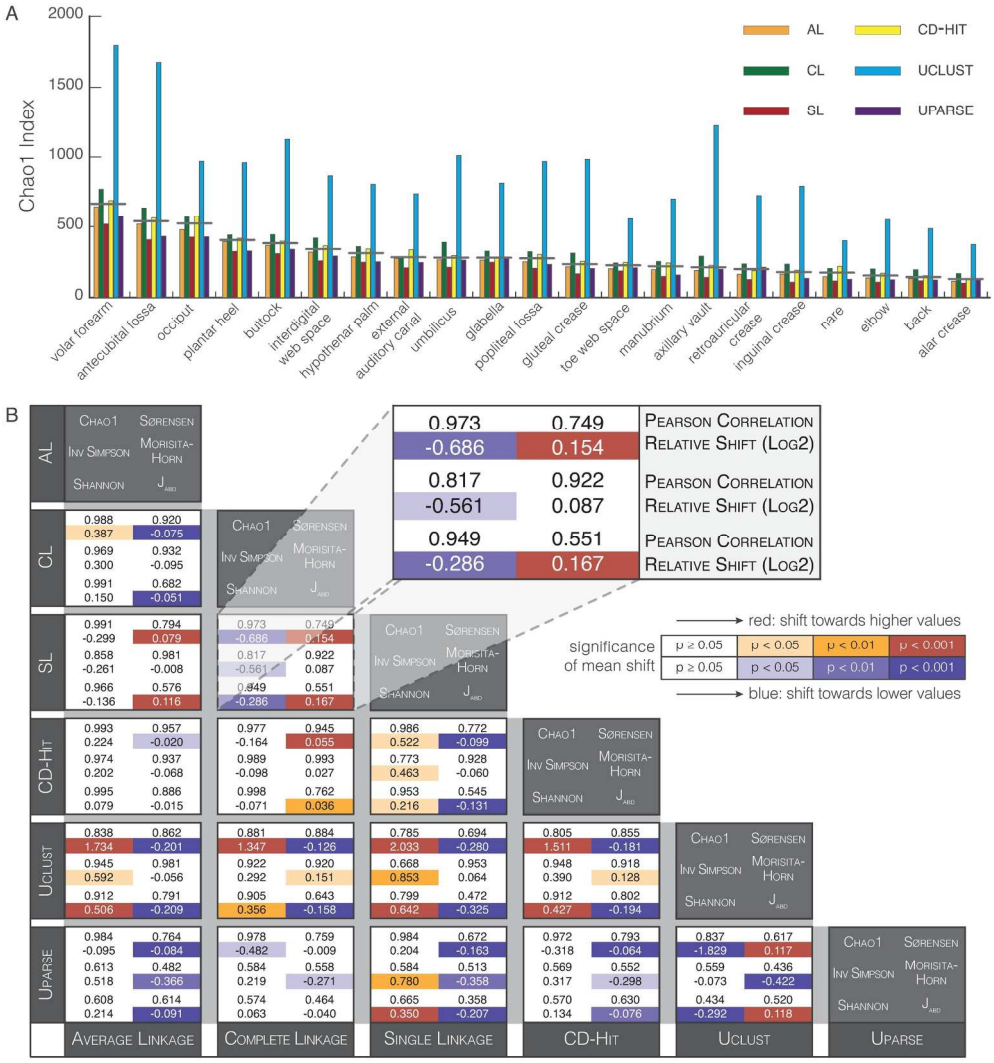


Figure 3. Clustering methods introduce differential bias to ecological data descriptions. (A) Chao1 richness estimates for the 21 samples of the HSM dataset when clustering to 97% sequence similarity using different methods. Vertical grey bars indicate median values. (B) Biases to estimators of local diversity (Chao1, inverse Simpson and Shannon indices) and community similarity (Sørensen-Dice-Czekanowski, Morisita-Horn and abundance-corrected Jaccard indices; see Text S1) for the same 21 skin samples. For every pair of clustering methods, Pearson correlations (upper cells) indicate differences in diversity ranking between samples per index. Average shifts in absolute diversity estimates per index ('relative shift', lower cells) between methods are calculated 'row-wise' (e.g., for the highlighted cell, read 'sl provided lower Chao1 estimates than cl'). Cell coloring indicates significance levels and direction of absolute shifts between methods (Mann-Whitney-U test). Raw data on diversity estimates is provided in Table S3.
189x201mm (300 x 300 DPI)

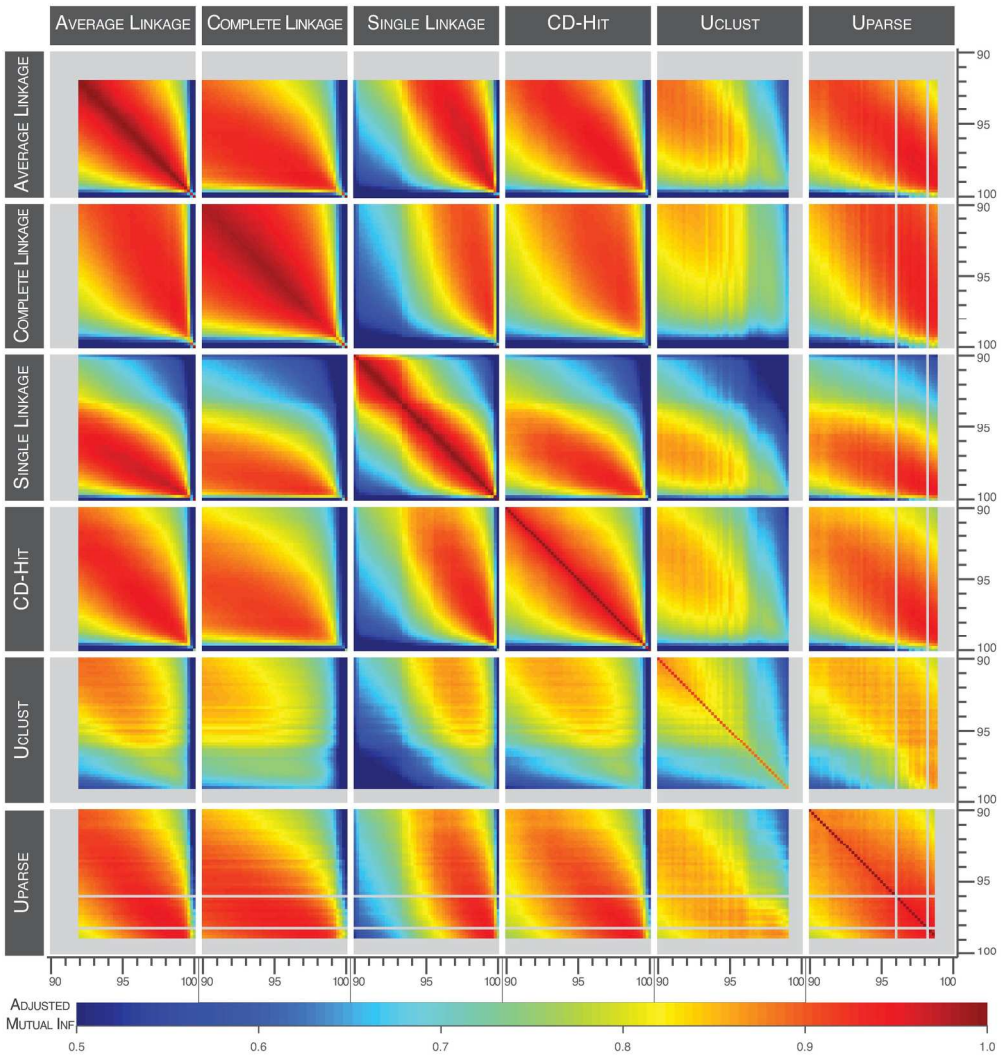


Figure 4. Differences in cluster composition between methods across wide threshold ranges. A global dataset of 887,870 16S sequences was clustered to thresholds ranging from 90-100% sequence similarity, in steps of 0.2% (corresponding to ~2.6 differences across the full sequence length). Pairwise partition similarities between methods and across thresholds were calculated as Adjusted Mutual Information (AMI). To calculate partition similarities of methods 'against themselves' (subplots on diagonal), clustering was re-run with randomized order of sequences. Note that algorithm memory requirements prohibited al clustering of the full set to <92% similarity and uclust/uparse clustering to >99% similarity; moreover uparse consistently crashed when clustering the dataset to 96.0% or 98.2% similarity (grey lines in the corresponding plots). Equivalent plots on NMI and ARI similarities, and for the HSM dataset, are provided as Figures S1-5. Raw data on partition similarities provided in Tables S4-9.
187x197mm (300 x 300 DPI)

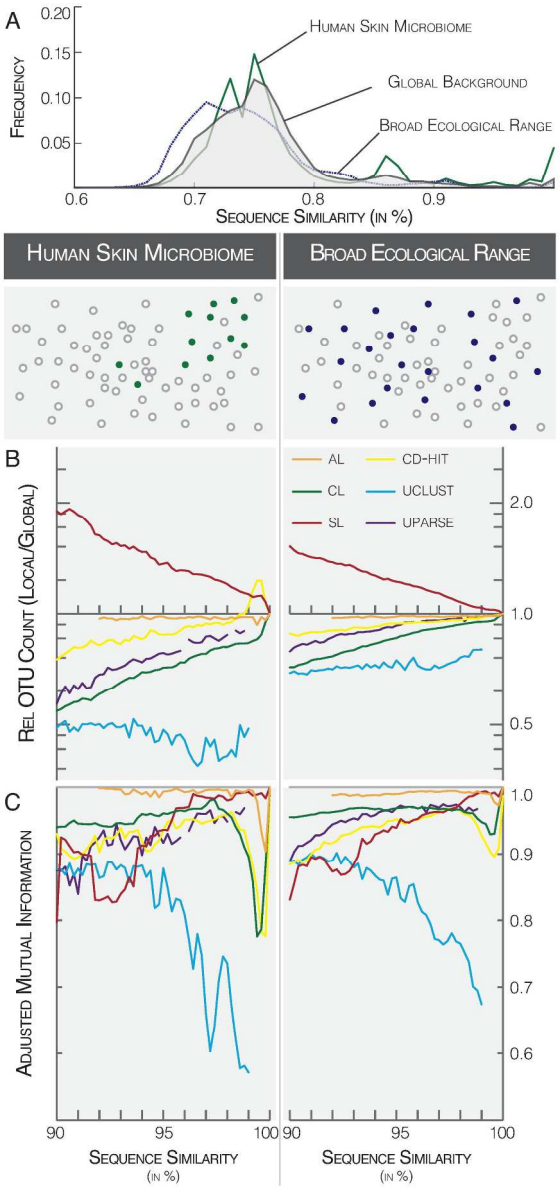


Figure 5. Robustness to clustering context. (A) The HSM and an artificially generated dataset of broad ecological range (BER, see Table S1) were extracted as 'local' subsets from the global set of 887,870 16S sequences. Pairwise internal sequence similarities were calculated based on 10 randomly drawn sets of 10,000 sequences per dataset. Compared to the global background, internal sequence similarities were significantly higher for the HSM set, and significantly lower for the BER set ($p << 10\text{-}16$, Mann-Whitney-U test). This corresponds to the patterns of filled dots (HSM/BER) over circles (global background) in the lower panel in (A). (B) Relative OTU counts when clustering HSM and BER sets in the presence ('global context') and absence ('local context') of the full global sequence space. (C) Partition similarities between local and global context, expressed as Adjusted Mutual Information (AMI).
180x385mm (300 x 300 DPI)

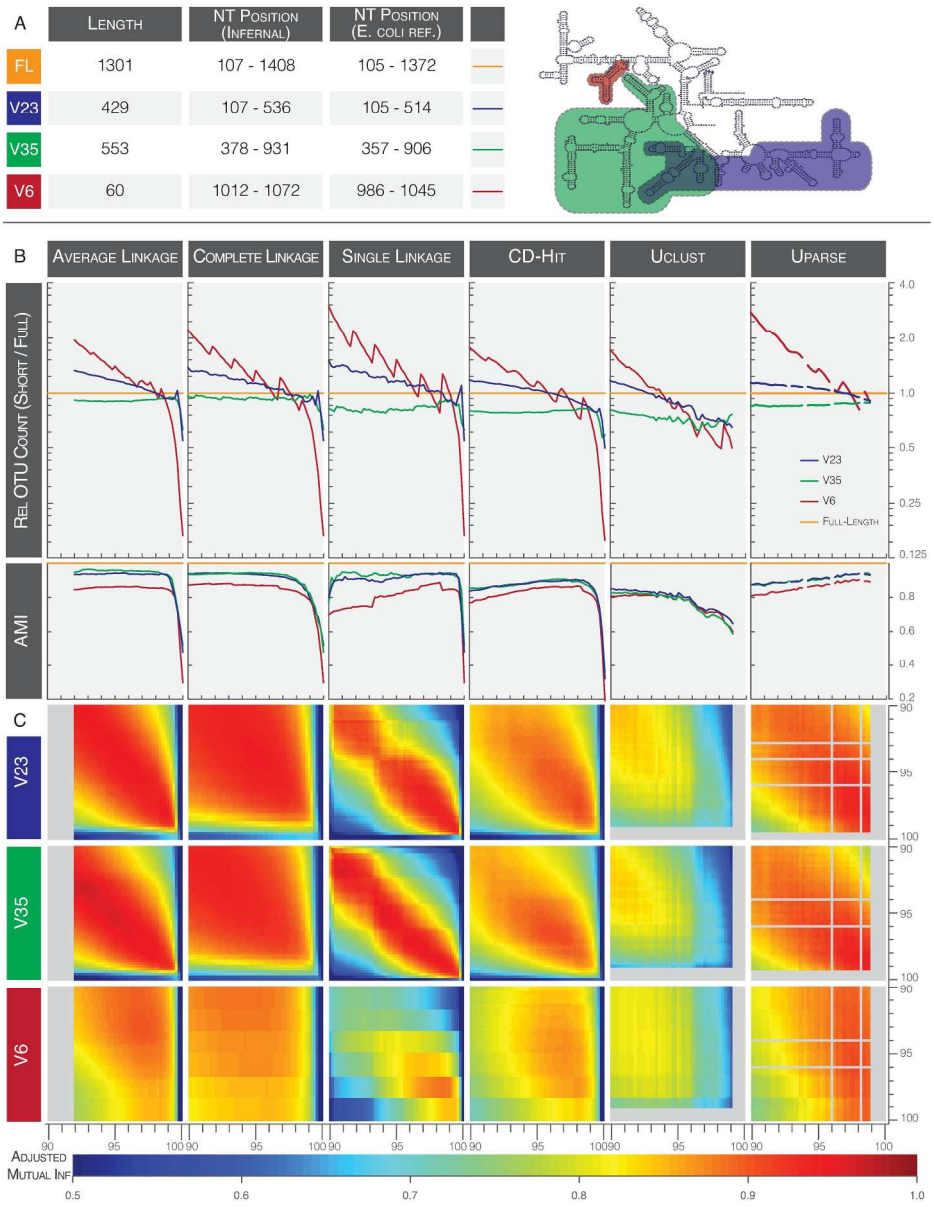


Figure 6. Robustness to the choice of SSU gene subregion. (A) Extraction of selected hypervariable subregions from near full-length 16S sequence alignments ('FL'). Sequence length (left column) and nucleotide positions in the ssu-align bacterial 16S model (middle column, Nawrocki, 2009), and the E. coli reference 16S sequence (right column) for subregionsV23, V35 and V6 were chosen following Schloss (2010), and indicated in the secondary structure resolved E. coli 16S sequence on the right (modified from an image kindly provided by Harry Noller, University of California, Santa Cruz). (B) Relative OTU counts of subregion clustering over full-length clustering and partition similarities as Adjusted Mutual Information (AMI) when clustering to the same nominal similarity threshold according to different methods. (C) Partition similarities across clustering thresholds reveal differential trends in robustness between methods.
230x297mm (300 x 300 DPI)

	al		cl		sl		cd-hit		uclust	
	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel
al										
cl	0.938 0.942	0.916 0.261								
sl	0.943 0.945	0.936 -0.241	0.885 0.888	0.852 -0.502						
cdhit	0.957 0.960	0.949 0.112	0.930 0.936	0.931 -0.149	0.915 0.917	0.888 0.352				
uclust	0.842 0.854	0.816 1.320	0.809 0.834	0.793 1.058	0.800 0.807	0.755 1.560	0.818 0.837	0.787 1.208		
uparse	0.963 0.965	0.870 -0.207	0.948 0.951	0.902 -0.469	0.910 0.911	0.711 0.033	0.945 0.948	0.868 -0.319	0.839 0.851	0.748 -1.528

	Inv_Simpson	Shannon	Chao1	Soerensen-Dice	Morisita_Horn	Jacc_abd
	Pearson Cor Shifts	Pearson Cor Shifts	Pearson Cor Shifts	Pearson Cor Shifts	Pearson Cor Shifts	Pearson Cor Shifts
OC	0.687 0.631	0.762 0.787	0.886 0.986	0.740 0.771	0.579 -0.411	0.669 0.712
ARI	0.802 0.770	0.861 0.910	0.861 0.814	0.837 0.882	0.767 0.688	0.757 0.893
AMI	0.543 0.673	0.611 0.891	0.884 0.866	0.642 0.864	0.493 0.381	0.606 0.858
NMI	0.541 0.671	0.610 0.895	0.885 0.870	0.648 0.870	0.487 0.383	0.611 0.864

	al	cl	sl	cdhit	uclust	uparse
	AMI NMI ARI	AMI NMI ARI	AMI NMI ARI	AMI NMI ARI	AMI NMI ARI	AMI NMI ARI
al						
cl	0.607 0.575 0.234					
sl	0.869 0.826 0.748	0.687 0.608 0.574				
cdhit	0.945 0.944 0.858	0.573 0.535 0.267	0.830 0.755 0.628			
uclust	0.716 0.820 0.363	0.626 0.794 0.626	0.766 0.668 0.612	0.725 0.768 0.407		
uparse	0.884 0.852 0.147	0.866 0.857 0.702	0.310 0.215 0.012	0.705 0.643 0.203	0.936 0.922 0.645	

	al		cl		sl		cd-hit		uclust	
	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel	AMI NMI	ARI OC_rel
al										
cl	0.915 0.853	0.479 0.738								
sl	0.806 0.599	0.481 0.399	0.444 0.029	-0.271 0.661						
cd-hit	0.993 0.991	0.960 -0.017	0.899 0.830	0.488 0.252	0.743 0.498	0.376 0.582				
uclust	0.840 0.777	0.190 -0.417	0.892 0.924	0.595 -0.103	0.310 -0.043	-0.586 0.219	0.829 0.790	0.350 0.633		
uparse	0.918 0.965	0.760 -0.664	0.813 0.825	0.539 -0.383	0.668 0.447	0.110 -0.057	0.934 0.976	0.826 0.361	0.754 0.811	0.465 0.736

Sequence Data and Preprocessing

We obtained a global, comprehensive dataset of publicly available full-length bacterial 16S sequences as described previously (Schmidt et al., 2014). In short, SSU sequences were downloaded from NCBI GenBank[1] (Benson et al., 2013) and from the genomes available in the NCBI Reference Sequence Database[2] (RefSeq, Pruitt et al., 2011) and pre-filtered for annotations as 'ribosomal RNA' or 'rRNA' and for a minimum length of 1,000bp. Sequences were aligned to a bacterial 16S consensus model (provided in the package ssu-align, Nawrocki, 2009) using the secondary structure-aware *Infernal* aligner (Nawrocki et al., 2009). Sequences which provided negative *Infernal* alignment scores, or aligned with higher confidence to an archaeal 16S or eukaryal 18S model, were removed from the dataset. To obtain an alignment of uniform length, comprising the same amount of information per sequence, all sequences were pruned at manually chosen conserved flanking positions (alignment position 107 to 1,408, yielding a total length of 1,301 alignment columns). We removed an additional 19.7% of sequences which were flagged as chimeric by *uchime* (Edgar et al., 2011), run with a set of reference sequences generated *de novo* from the entire alignments. After these pre-processing steps, the dataset used in this study comprised 887,870 bacterial 16S sequences of which 673,128 or 75.8% were unique; this dataset is referred to as 'global' set in the main text.

From this set, two 'local' subsets were extracted for further analysis: (i) the well-characterized *human skin microbiome* dataset (HSM, Grice et al., 2009), comprising 90,620 sequences after filtering steps; and (ii) an artificially generated dataset of *broad ecological range* (BER), comprising 53,999 sequences from 18 distinct studies focusing on distinct environments (see Table S1 for further details). Moreover, three global datasets of simulated 'short reads' were generated by extracting subregions V23, V35 and V6 from the full alignments (details in main text, and in particular in Figure 6).

Sequence Clustering into Operational Taxonomic Units

Sequences were clustered into OTUs using six different methods: *average linkage* (*al*), *complete linkage* (*cl*) and *single linkage* (*sl*) hierarchical clustering, as well as heuristic *cd-hit*, *uclust* and *uparse* clustering. For each method, OTUs were clustered to sequence similarity thresholds of 90-100%, in steps of 0.2%. For the full global set of 887,870 bacterial 16S sequences, clustering to the full threshold range was not possible for some methods due to prohibitive memory requirements of the algorithms: (i) hierarchical *al* clustering was only performed to ≥92% sequence similarity; (ii) *uclust* and *uparse* clustering was only performed to <99% sequence similarity, as higher similarity thresholds required ≥4 gigabytes of RAM, which is the limit for the freely available 32bit versions of these tools; (iii) clustering was (reproducibly) unsuccessful for *uparse* at additional individual datapoints, depending on the dataset (e.g., the full global set was not clustered to 96.0% and 98.2% similarity).

We were not able to cluster the full global 16S datasets, or subsets thereof (≥100,000 sequences) using *ESPRIT* (Sun et al., 2009), *ESPRIT-Tree* (Cai and Sun, 2011) and *mothur* (Schloss et al., 2009), even when providing excessive computational resources (running on a multicore machine with 48 CPUs and 1 terabyte of RAM). This is likely due to the computationally expensive calculation of a pairwise sequence distance matrix. However, it has been shown that hierarchical clustering as implemented in *hpc-clust* (used in this study) and *mothur* provides virtually identical partitions of the data (Matias Rodrigues and von Mering, 2014). Moreover, *ESPRIT*

[1] http://www.ncbi.nlm.nih.gov/genbank/, accessed in April 2012

[2] http://www.ncbi.nlm.nih.gov/RefSeq/, accessed in March 2012

and *ESPRIT-Tree* are (slightly heuristic) implementations of the *cl* and *al* algorithms, albeit using a pairwise alignment strategy, rather than multiple sequence alignments. Thus, we are confident that our findings on hierarchical clustering algorithms may be portable to *ESPRIT*, *ESPRIT-Tree* and *mothur*, although we have not explicitly tested this.

In the following, command line options for running the different clustering software tools (as pseudo-code) are given.

HPC-Clust

Hierarchical *al, cl* and *sl* clustering to the full threshold range were implemented in single runs on a 256 CPU computer cluster using *openmpi*[3].

```
#Run hpc-clust
> hpc-clust-mpi -al true -cl true -sl true -t 0.8 --dfunc gap -ofile ${clustering_merge_file} ${alignment_file}
#Parameters:
#"-t"        =>    minimum similarity threshold
#"--dfunc"   =>    sequence distance calculator. "gap" stands for the "onegap"
#                  calculator which counts gaps of any length as single mismatch
#"-ofile"    =>    output file (records merges of clusters along thresholds)
#
#Demarcate OTUs from merge files
> make-otus.sh ${alignment_file} ${clustering_merge_file} ${threshold}
#"threshold" specifies the similarity threshold to which OTUs are being demarcated.
```

CD-HIT

Our protocol for *cd-hit* clustering was modeled to be consistent with the *cd-hit-otu* pipeline suggested by Li et al (2012). However, to control for potential variability in sequence preprocessing, we used the same set of preprocessed, *uchime*-filtered sequences as input for *cd-hit* as for the hierarchical methods and *uclust*. Moreover, we used word lengths (k-mer sizes) of 11; shorter word lengths were tested, but these did not provide significantly different partitions, while clustering took longer to compute. We used *cd-hit* in version 4.5.4, build 2012-08-25. All runs were performed on a multicore computer (48 CPUs, 1 terabyte of RAM).

```
#Run cd-hit
> cd-hit-est -i ${unaligned_sequences_file} -o ${output_file} -c ${threshold} -T ${no_of_cpus} -M 100000 -n 11 -d 150
#Parameters
#"-T"        =>    number of cores when running in parallel
#"-M"        =>    memory threshold (in megabytes)
#"-n"        =>    word length
#"-d"        =>    length of description in output file
```

[3] 'message passing library', http://www.open-mpi.org

Uclust

We performed *uclust*[4] (Edgar, 2010) clustering on the *uchime*-filtered sets of sequences (see above), but using unaligned sequences as input.

```
#Run uclust
> usearch -cluster_fast ${unaligned_sequences_file} -id ${threshold}
```

Uparse

As *uparse* implements on-the-fly filtering for chimeric sequences, we performed *uparse*[5] (Edgar, 2013) runs on full, non-chimera-filtered sets of unaligned sequences which were subsequently mapped to the *uchime*-filtered sets used for the other methods. Moreover, Edgar (2013) suggests to remove singleton OTUs (containing only one sequence) as 'spurious' by default; however, since our dataset consisted of near full-length, high-quality reads, and since other methods do not implement cluster size-filtering, we did not remove singleton clusters from *uparse* partitions. As for *cd-hit* and *uclust*, *uparse* runs were performed on a multicore computer (48 CPUs, 1 terabyte of RAM).

```
#Dereplicate dataset
> usearch -derep_fulllength ${unaligned_sequences_file} -output ${unaligned_dereplicated} -sizeout
#Sort by size, but do not discard singletons
> usearch -sortbysize ${unaligned_dereplicated} -output ${unaligned_sorted} -minsize 1
#Run uparse
> usearch -cluster_otus ${unaligned_sorted} -otuid ${threshold} -otus ${representatives}
#Run uchime on cluster representatives
> usearch -uchime_ref ${representatives -db ${reference_database} -strand plus -nonchimeras ${representatives_nonchimeric}
#Map reads back to (nonchimeric) representatives
> usearch -usearch_global ${unaligned_sequences_file} -db ${representatives_nonchimeric} -id ${threshold} -strand plus -uc ${output_file}
```

[4] version 6.0.307

[5] version 7.0.1001_i86linux32

Indices of Community Diversity

Local community diversity ('α-diversity')

For the 21 samples of the HSM dataset, we estimated local community diversity based on three widely used indices. The *Chao1* richness estimator (Chao, 1984) was calculated as follows:

$$S_{Chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)}$$

where S_{obs} is the observed richness (number of OTUs) and n_1 and n_2 are the number of *singleton* (only one sequence) and *doubleton* (two sequences) OTUs, respectively. In other words, the Chao1 index corrects for the number of *unseen* OTUs based on the number of *rare* OTUs.

The *Shannon* index (Shannon, 1948) was calculated as follows:

$$H = -\sum_i \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right)$$

where n is the total number of sequences and n_i is the size of class i. In other words, the Shannon index is formulated as an *entropy,* describing the uncertainty when determining the OTU membership of a given sequence in the sample.

Finally, the *inverse Simpson* index (Simpson, 1949) was calculated as follows:

$$ISI = \sum_i \frac{n_i(n_i - 1)}{n(n-1)}$$

Thus defined, the inverse Simpson index is the probability that two sequences randomly drawn from a sample belong to the same OTU.

Community similarity ('β-diversity')

We calculated pairwise community similarity between the 21 HSM samples according to three widely used indices. We calculated the abundance-based Jaccard index, with a correction for raw / unseen taxa as suggested by Chao et al (2004):

$$J_{abd}(A,B) = \frac{U_{est} V_{est}}{U_{est} + V_{est} - U_{est} V_{est}}$$

where U_{est} and V_{est} are the (unseen taxa-corrected) estimates of total relative abundances of shared species in groups A (U_{est}) and B (V_{est}), defined as:

$$U_{est} = \sum_i^{S_{A,B}} \frac{a_i}{n_A} + \frac{n_B - 1}{n_B} \frac{f_{+1}}{2f_{+2}} \sum_i^{S_{A,B}} \frac{a_i}{n_A} I(b_i = 1)$$

$$V_{est} = \sum_i^{S_{A,B}} \frac{b_i}{n_B} + \frac{n_A - 1}{n_A} \frac{f_{1+}}{2f_{2+}} \sum_i^{S_{A,B}} \frac{b_i}{n_B} I(a_i = 1)$$

where $S_{A,B}$ is the number of shared OTUs between groups A and B, a_i is the size of OTU i in A, b_i the size of OTU i in B, n_A and n_B are the total number of sequences in A and B. *I(expression)* is an indicator function,

defined as I = 1 if 'expression' is true and I = 0 otherwise. Finally, f_{+1} and f_{+2} are the number of shared OTUs that are singletons and doubletons in partition A, while f_{1+} and f_{2+} are the number of shared OTUs that are singletons and doubletons in partition B. Thus, the number of 'unseen shared taxa' is estimated based on the number of 'observed shared taxa' between the partitions. In the above formulation, the abundance-corrected Jaccard index is defined as community *similarity*, so that J = 1 describes perfectly identical communities, while J = 0 if no taxa are shared. Note that the abundance-corrected Jaccard index is implemented as "*jabund*" in the *mothur* suite[6].

We calculated the *Sørensen-Dice-Czekanowski* coefficient (Dice, 1945) in the raw abundance-based version as defined by Chao et al (2004):

$$SDC = \frac{2UV}{U+V} = \frac{2\sum_{i}^{S_{A,B}} \frac{a_i}{n_A} \sum_{i}^{S_{A,B}} \frac{b_i}{n_B}}{\sum_{i}^{S_{A,B}} \frac{a_i}{n_A} + \sum_{i}^{S_{A,B}} \frac{b_i}{n_B}}$$

where *U* and *V* are the sums of relative abundances of individuals in shared taxa in groups A and B. In the above formulation, the SDC is defined as an index of community *similarity*; it is closely related to the widely used *Bray-Curtis dissimilarity* index (Bray and Curtis, 1957).

The abundance-based *Morisita-Horn* overlap index (Horn, 1966) was calculated as follows:

$$MH = \frac{2\sum_{i}^{S} a_i b_i}{\left(\sum_{i}^{S} \frac{a_i^2}{n_A^2} + \sum_{i}^{S} \frac{b_i^2}{n_B^2}\right) n_A^2 n_B^2}$$

where *S* is the total number of unique OTUs between groups, n_A and n_B are the total number of sequences in groups A and B, and a_i and b_i are the absolute frequencies of taxon *i* in A and B. Values for MH range between 0 (no overlap between communities) and 1 (all taxa are present in both groups in equal abundances).

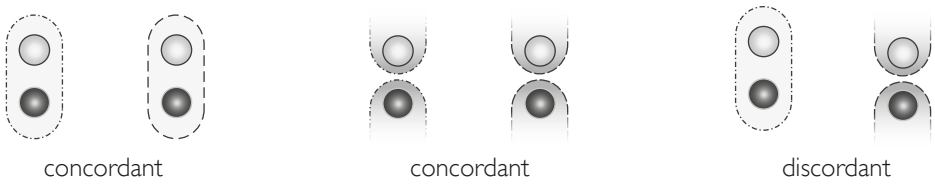
[6] http://www.mothur.org/wiki/Jabund

Assessing Partition Similarity

Similarity in cluster composition was calculated using the pair counting-based *Adjusted Rand Index* (ARI) and the information theoretic-based *Normalized Mutual Information* (NMI) and *Adjusted Mutual Information* (AMI).

Adjusted Rand Index (ARI)

As indicated already by name, pair counting-based indices quantify the similarity between two partitions by counting individual pairs of sequences as either *concordant* or *discordant*. A pair of sequences is concordant across partitions if it either clusters together in both partitions (*'agree to agree'*) or does not cluster together in either partition (*'agree to disagree'*). In contrast, discordant sequence pairs cluster together in one partition, but into different OTUs in the other.



Concordant and discordant pairs of sequences.

The *Rand Index* (Rand, 1971) of partition similarity weighs counts of concordant and discordant pairs of sequences as follows:

$$RI = N_{concordant} / \binom{n}{2}$$

where n is the total number of sequences and $N_{concordant}$ is the number of concordant pairs. In other words, the Rand Index is the ratio of concordant pairs per total pairs. Based on the observation that the Rand Index does not take a constant expected value between random partitions, Hubert and Arabie (1985) proposed an adjusted form which corrects for chance based on a hypergeometric randomness model. The *Adjusted Rand Index* (ARI) is calculated as follows:

$$ARI = \frac{Index - Expected\_Index}{Max\_Index - Expected\_Index} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where a_i is the size of OTU i in partition A, b_j is the size of OTU j in partition B and $n_{i,j}$ is the number of sequences clustering into OTU i in partition A and OTU j in partition B (i.e., the i,j-th entry in the contingency table between partitions). ARI values range between -1 (complete discordancy – sequences grouping together in A never group together in B) to 1 (perfectly identical partitions); ARI = 0 indicates random similarity as expected based on cluster size distributions.

Arguably, one inherent drawback of pair counting-based measures is the dominance of large clusters. Since the number of pairwise comparisons scales quadratically with sequence count, large clusters will contribute disproportionally more to the similarity / dissimilarity signal than smaller clusters. In an extreme case, merging two large clusters X_A and Y_A in partition A into one cluster Z_B in partition B will provide a large number of

discordant sequence pairs, although both partitions remain highly similar from a set point of view. Nevertheless, pair counting-based indices are also appealingly intuitive: they provide an immediate notion of relative partition concordance at the level of pairs of sequences.

Information theoretic-based indices (NMI, AMI)

More recently, information theoretic-based indices have received increasing attention in the clustering literature, not least due to their strong theoretical background (Vinh et al., 2009). Consider a partition A of *i* clusters of sizes a_i. The *entropy* of partition A quantifies the uncertainty when determining a given sequence's cluster membership in A; it is calculated as follows:

$$H(A) = -\sum_i \frac{a_i}{n} \log\left(\frac{a_i}{n}\right)$$

H(A) = 0 indicates the 'singleton partition', i.e. a partition with only one cluster in which there is no uncertainty about cluster membership. The information overlap between two partitions A and B can be expressed based on these partitions' entropies, as the *Mutual Information* (MI):

$$I(A,B) = \sum_i \sum_j n_{i,j} \log\left(\frac{n_{i,j} n}{a_i b_j}\right)$$

In other words, *I(A,B)* quantifies the mutual dependence between partitions A and B and "measures how much knowing one of these [partitions] reduces our uncertainty about the other" (Vinh et al., 2009). As I(A,B) is upper bounded by the entropies H(A) and H(B), several mathematically related or even equivalent normalizations have been proposed, such as the *Variation of Information* (VI) by Meilă (2005) or different versions of the *Normalized Mutual Information* (NMI). As defined by Fred and Jain (2003), the NMI is calculated as follows:

$$NMI = \frac{-2I(A,B)}{H(A)+H(B)} = \frac{-2\sum_i \sum_j n_{i,j} \log\left(\frac{n_{ij} n}{a_i b_j}\right)}{\sum_i a_i \cdot \log\left(\frac{a_i}{n}\right) + \sum_j b_j \log\left(\frac{b_j}{n}\right)}$$

NMI values range between 0 (no shared information between partitions) to 1 (perfectly identical partitions). In the context of OTU demarcation, both NMI and VI have previously been used to test the agreement of OTU sets with differently defined taxonomic ground truth partitions (White et al., 2010; Cai and Sun, 2011; Sun et al., 2011; Bonder et al., 2012; Wang et al., 2013). Noting that variations in cluster counts cause systematically shifting NMI baseline values, Vinh et al. (2009) proposed the *Adjusted Mutual Information* (AMI) measure which uses a hypergeometric permutation model to correct for these effects:

$$AMI = \frac{I(A,B) - E\{I(M) \mid a,b\}}{\sqrt{H(A)H(B)} - E\{I(M) \mid a,b\}}$$

where $E\{I(M)|a,b\}$ is the expected average Mutual Information for all theoretically possible contingency tables with marginals *a* and *b*; in other words, $E\{I(M)|a,b\}$ is the expected Mutual Information for the observed distributions of cluster sizes in partitions A and B. It is defined as

$$E\{I(M)\,|\,a,b\} = \sum_i \sum_j \sum_{n_{i,j}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{i,j}}{n} \log\left(\frac{n_{i,j}n}{a_i b_j}\right) \frac{a_i!\,b_j!\,(n-a_i)!\,(n-b_j)!}{n!\,n_{i,j}!\,(a_i-n_{i,j})!\,(b_j-n_{i,j})!\,(n-a_i-b_j+n_{i,j})!}$$

Vinh et al could show in simulation studies that AMI values do not suffer from a systematically shifting baseline with shifting cluster counts. Similarly to ARI, values for AMI range between [-1, 1]; AMI = 1 describes perfectly identical partitions, AMI = 0 indicates 'random' shared information as expected by chance for two partitions of the given cluster size distributions. We used both NMI and AMI to assess partition similarity across clustering methods, and for varying clustering parameters.

References

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013) GenBank. *Nucleic Acids Research* **41**: D36–42.

Bonder, M.J., Abeln, S., Zaura, E., and Brandt, B.W. (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics (Oxford, England)* **28**: 2891–2897.

Bray, J.R. and Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27**: 325–349.

Cai, Y. and Sun, Y. (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research* **39**: e95.

Chao, A. (1984) Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* **11**: 265–270.

Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.-J. (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8**: 148–159.

Dice, L.R. (1945) Measures of the Amount of Ecologic Association Between Species. *Ecology* **26**: 297–302.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)* **26**: 2460–2461.

Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Publishing Group* **10**: 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)* **27**: 2194–2200.

Fred, A.L.N. and Jain, A.K. (2003) Robust Data Clustering. pp. 128–136.

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., et al. (2009) Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324**: 1190–1192.

Horn, H.S. (1966) Measurement of" overlap" in comparative ecological studies. *American Naturalist* 419–424.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification* **2**: 193–218.

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics* **13**: 656–668.

Matias Rodrigues, J.F. and von Mering, C. (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics (Oxford, England)* **30**: 287–288.

Meilă, M. (2005) Comparing Clusterings: An Axiomatic View. ACM, New York, NY, USA, pp. 577–584.

Nawrocki, E.P. (2009) Structural RNA Homology Search and Alignment Using Covariance Models. 1–281.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)* **25**: 1335–1337.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research* **40**: D130–D135.

Rand, W.M. (1971) Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**: 846–850.

Schloss, P.D., Westcott, S.L., Rabyn, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75**: 7537.

Schmidt, T.S.B., Matias Rodrigues, J.F., and von Mering, C. (2014) Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale. *PLOS Computational biology* **10**: e1003594.

Shannon, C.E. (1948) A Mathematical Theory of Communication. *At&T Tech J* **27**: 623–656.

Simpson, E.H. (1949) Measurement of Diversity. *Nature* **163**: 688–688.

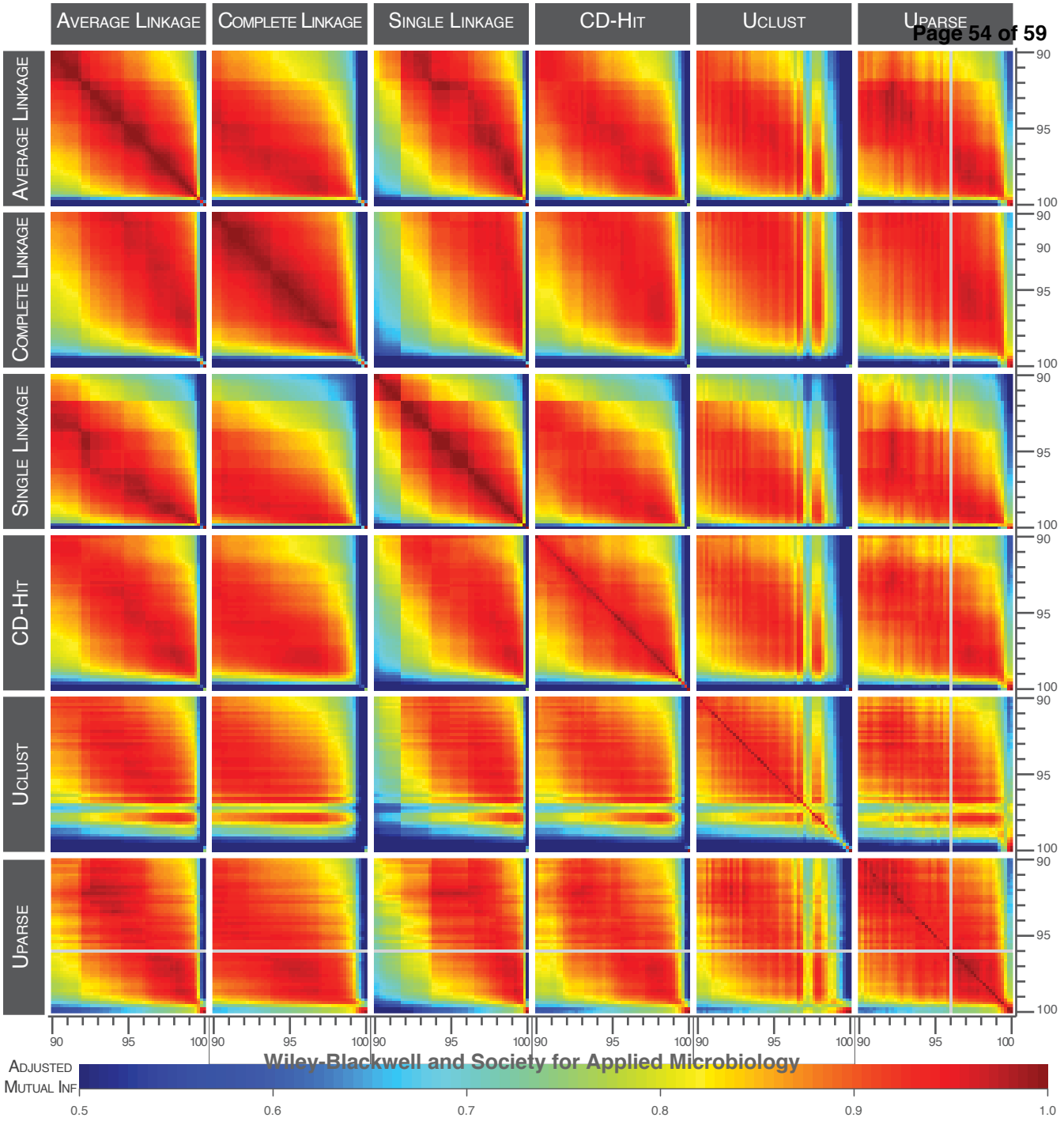
Sun, Y., Cai, Y., Huse, S.M., Knight, R., Farmerie, W.G., Wang, X., and Mai, V. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in bioinformatics* **13**: 107–121.

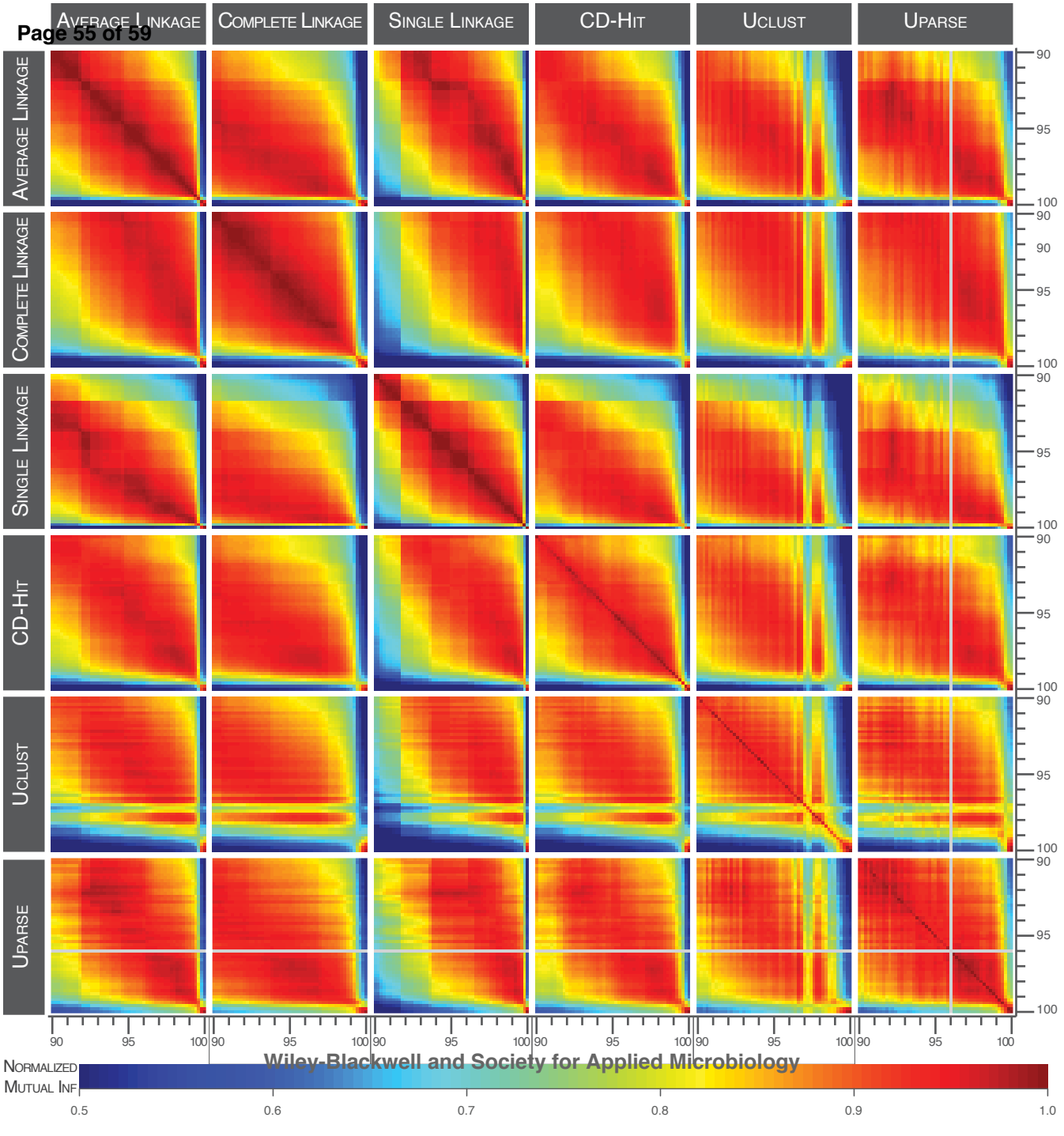
Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W., and Farmerie, W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research* **37**: e76–e76.

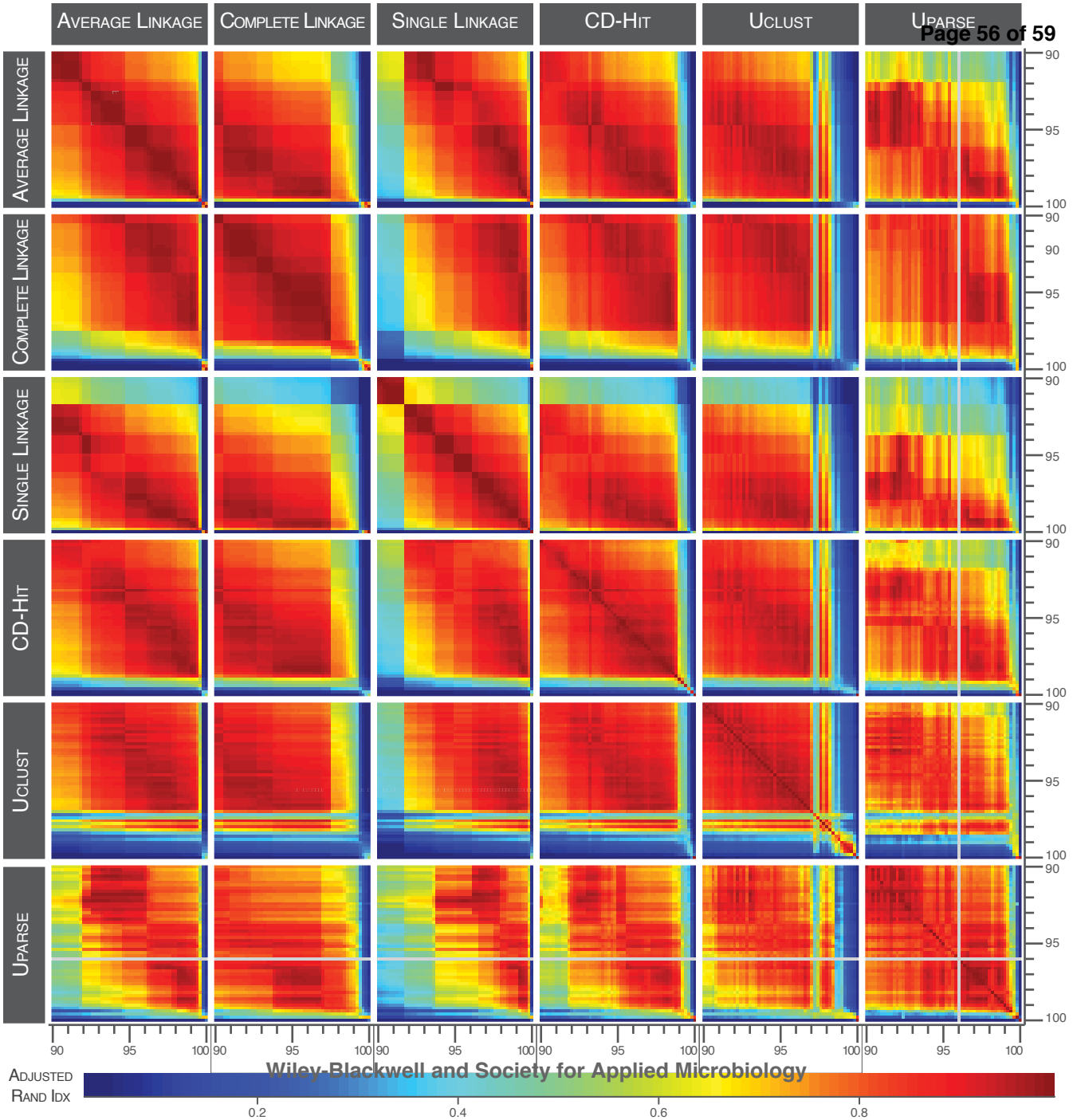
Vinh, N.X., Epps, J., and Bailey, J. (2009) Information theoretic measures for clusterings comparison. ACM Press, New York, NY, pp. 1073–1080.

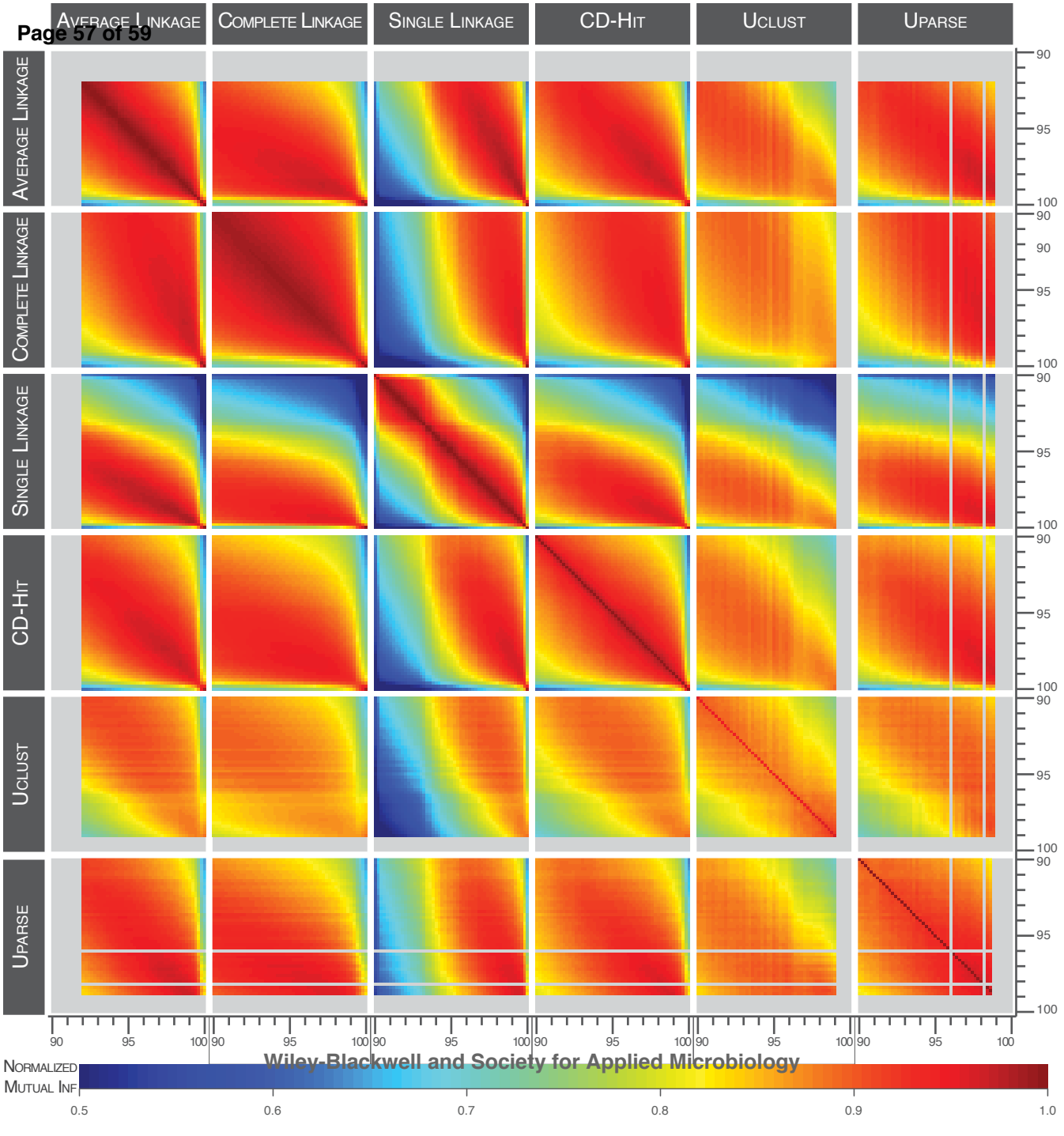
Wang, X., Yao, J., Sun, Y., and Mai, V. (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics* **14**: 43.

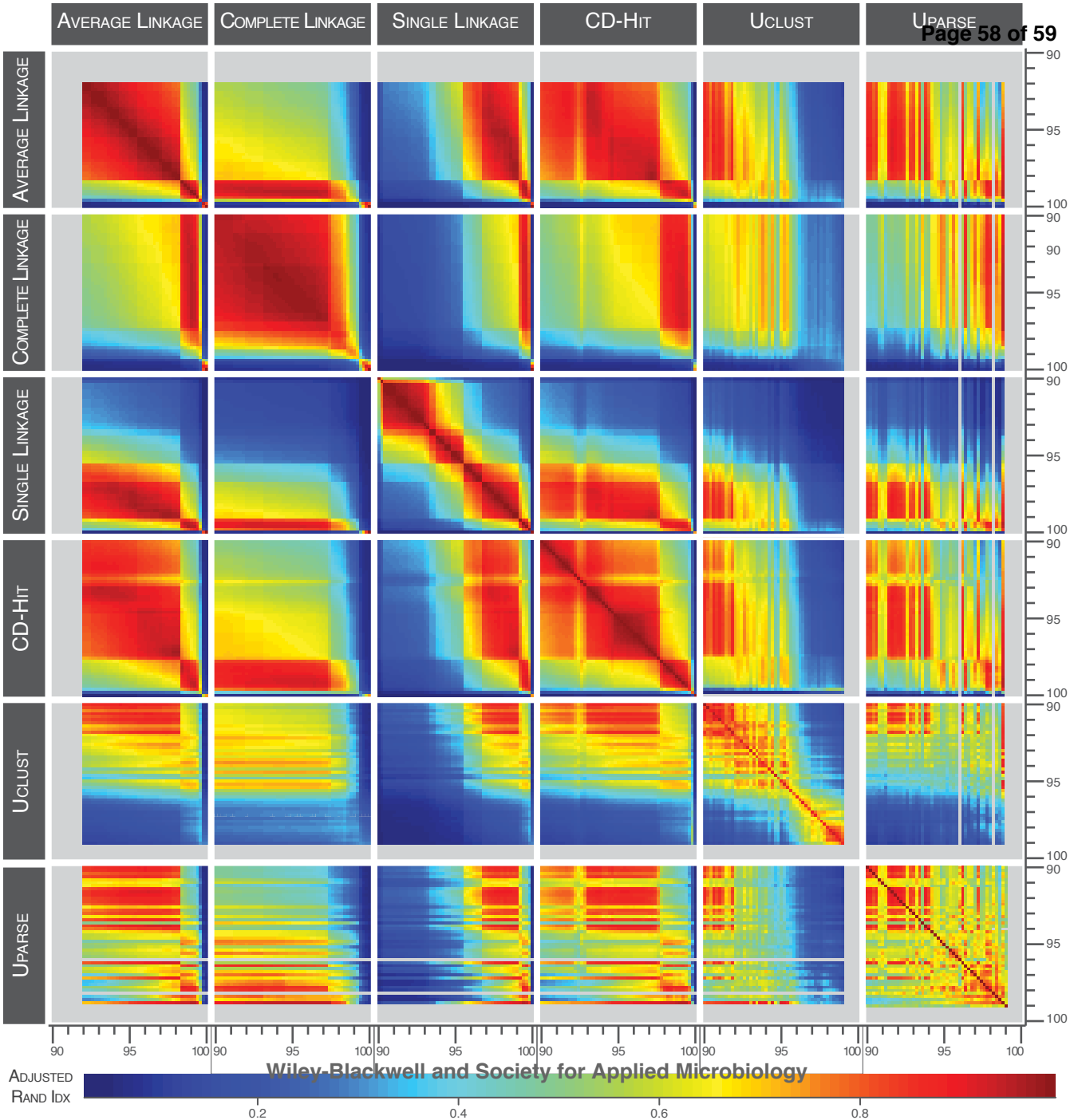
White, J.R., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M. (2010) Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics* **11**: 152.

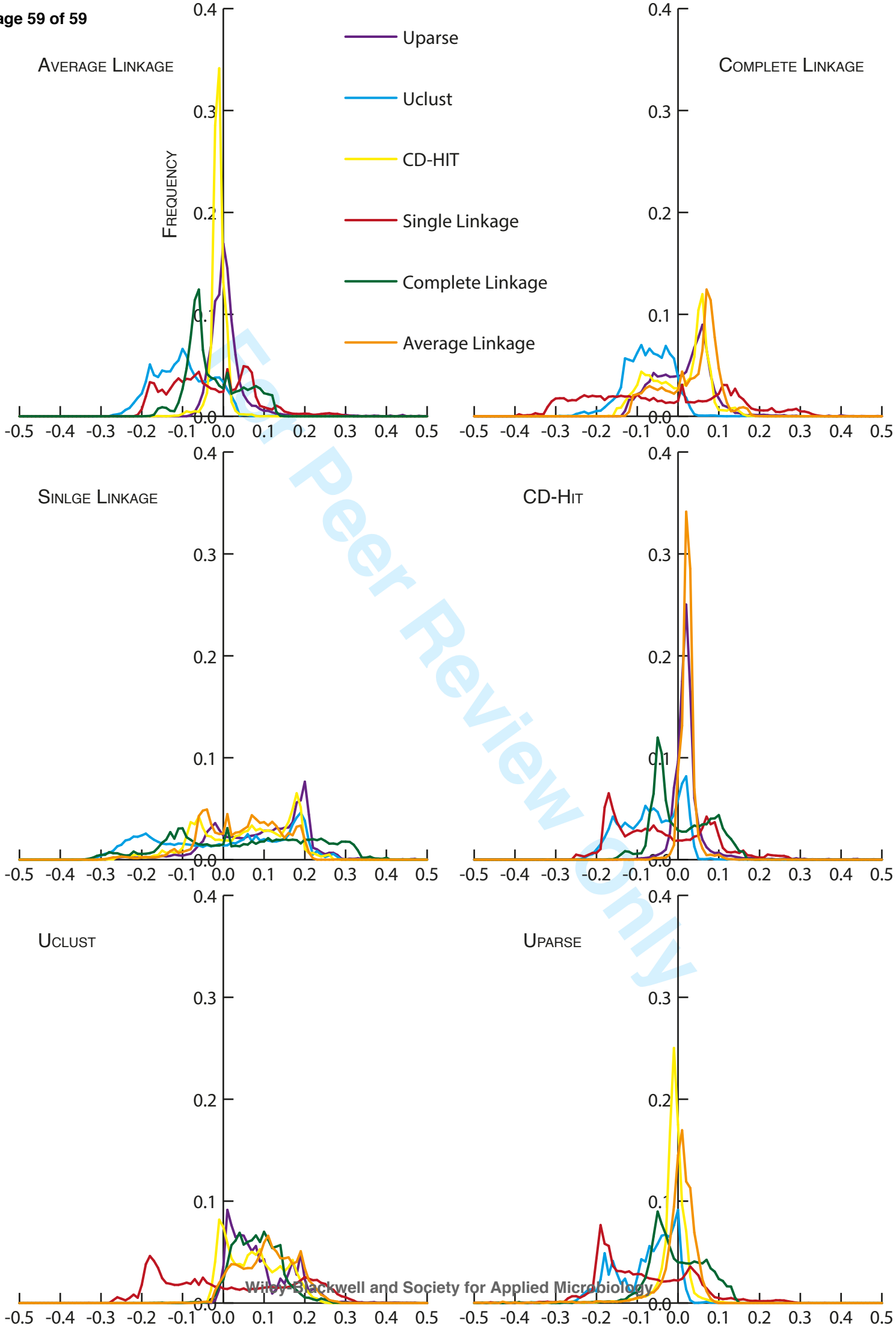












7. Selected Publications and Manuscripts

The main scientific output of this thesis is summarized in two first-author manuscripts:

Schmidt, T.S.B., Matias Rodrigues, J.F. & von Mering, C. (2014). **Limits to Robustness and Reproducibility in the Demarcation of Operational Taxonomic Units.** *under revision*

Contributions: TSBS, JFMR & CvM designed the study, TSBS conducted the experiments, TSBS, JFMR & CvM analyzed and interpreted the results, TSBS wrote the manuscript.

Schmidt, T.S.B., Matias Rodrigues, J.F. & von Mering, C. (2014). **Ecological Consistency of SSU rRNA-based Operational Taxonomic Units at a Global Scale.** PLOS Computational Biology, 10(4), e1003594. doi:10.1371/journal.pcbi.1003594

Contributions: TSBS, JFMR & CvM designed the study, TSBS conducted the experiments, TSBS, JFMR & CvM analyzed and interpreted the results, TSBS & CvM wrote the manuscript.

Moreover, several co-authored manuscripts contain contributions to completed and ongoing projects in context of this thesis:

Maier, L., Vyas, R., Cordova, C. D., Lindsay, H., Schmidt, T. S. B., Brugiroux, S., et al. (2013). **Microbiota-Derived Hydrogen Fuels *Salmonella Typhimurium* Invasion of the Gut Ecosystem.** Cell Host & Microbe, 14(6), 641–651. doi:10.1016/j.chom.2013.11.002

Contribution: TSBS analyzed a targeted 16S sequence dataset to confirm the taxonomic composition of the tested mice' gut microbiota.

Reprinted as appendix in section 8.1

Maier, S., Schmidt, T.S.B., Zheng, L., Peer, T., Wagner, V. & Grube, M. (2014). Specific Enrichment of Bacterial Communities in Lichens Forming Biological Soil Crusts. Biodiversity & Conservation, doi:10.1007/s10531-014-0719-1

Contribution: SM & TSBS analyzed the 16S sequence datasets.

Becker, E.*, Schmidt, T.S.B.*, Stanzel, C., Atrott, K., Biedermann, L., Rehman, A., Jonas, D., von Mering, C., Rogler, G., Frey-Wagner, I. (2014). **Influence of Isotretinoin Treatment on the Murine Gastrointestinal Tract.** *in preparation*

Contribution: EB & TSBS analyzed the 16S sequence data, in particular with respect to differences in community richness and composition between different groups of tested mice.

See section 8.4 for a full list of manuscripts and publications.

7.1 Limits to Robustness and Reproducibility in the Demarcation of Operational Taxonomic Units

Thomas Sebastian Benedikt Schmidt, João Frederico Matias Rodrigues & Christian von Mering (2014)

under revision

This manuscript is available in the printed version of this thesis, and will be available online as soon as permission by the journal publisher is obtained.

7.2 Ecological Consistency of SSU rRNA-based Operational Taxonomic Units at a Global Scale

Thomas Sebastian Benedikt Schmidt, João Frederico Matias Rodrigues & Christian von Mering (2014)

PLOS Computational Biology, 10(4), e1003594. doi:10.1371/journal.pcbi.1003594

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale

Thomas S. B. Schmidt, João F. Matias Rodrigues, Christian von Mering*

Institute for Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zürich, Switzerland

Abstract

Operational Taxonomic Units (OTUs), usually defined as clusters of similar 16S/18S rRNA sequences, are the most widely used basic diversity units in large-scale characterizations of microbial communities. However, it remains unclear how well the various proposed OTU clustering algorithms approximate 'true' microbial taxa. Here, we explore the ecological consistency of OTUs – based on the assumption that, like true microbial taxa, they should show measurable habitat preferences (niche conservatism). In a global and comprehensive survey of available microbial sequence data, we systematically parse sequence annotations to obtain broad ecological descriptions of sampling sites. Based on these, we observe that sequence-based microbial OTUs generally show high levels of ecological consistency. However, different OTU clustering methods result in marked differences in the strength of this signal. Assuming that ecological consistency can serve as an objective external benchmark for cluster quality, we conclude that hierarchical complete linkage clustering, which provided the most ecologically consistent partitions, should be the default choice for OTU clustering. To our knowledge, this is the first approach to assess cluster quality using an external, biologically meaningful parameter as a benchmark, on a global scale.

Citation: Schmidt TSB, Matias Rodrigues JF, von Mering C (2014) Ecological Consistency of SSU rRNA-Based Operational Taxonomic Units at a Global Scale. PLoS Comput Biol 10(4): e1003594. doi:10.1371/journal.pcbi.1003594

Editor: Jonathan A. Eisen, University of California Davis, United States of America

Received January 9, 2014; **Accepted** March 14, 2014; **Published** April 24, 2014

Funding: This work was supported by an ERC starting grant to CvM (UMICIS/242870), and by the Swiss National Science Foundation (31003A_135688). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mering@imls.uzh.ch

Introduction

Recent advances in sequencing technology have enabled researchers to characterize microbial diversity at previously unattainable scales. In large collaborative efforts, such as the Human Microbiome Project [1], selected environments have been probed to depths of millions of sequences, but even smaller-scale studies generate datasets of hundreds of thousands of reads. While providing great detail and resolution, datasets of such scopes pose a challenge to defining meaningful units of microbial diversity, and the choice of diversity unit definition may influence data analysis. Arguably, the gold standard for microbial diversity units are theory-informed definitions that would comply with a commonly accepted concept of bacterial speciation; in other words, operational units of diversity should approximate 'true' bacterial taxa [2]. This implies two frequently cited criteria for theory-compliant diversity units: they should reflect phylogeny (by representing *monophyletic* groups of organisms) and ecology, since ecological differentiation has been postulated as an important driver of bacterial speciation [2–8]. However, a unifying concept of bacterial speciation in fact remains controversial to the point of contesting the very existence of 'bacterial species' as such [2,9–11]. Nevertheless, approaches towards reconciling diversity unit definitions with evolutionary theory have received much attention. For example, the ecotype model of bacterial speciation defines basic diversity units as ecologically coherent groups of organisms whose diversity is confined by a cohesive genetic force [3,12], and dedicated algorithms have been developed to demarcate ecotypes from environmental sequencing data [4]. However, while ecotype simulation has been valuable in characterizing the diversity of selected environments [13], it has been noted that recognized diversity clusters within several microbial clades can conflict with ecotype theory [11,14].

Given the lack of a commonly accepted bacterial species concept, a phenomenological (pragmatic) approach to categorizing microbial diversity is often chosen in practice: *Operational Taxonomic Units* (OTUs), defined as clusters of 16S/18S small subunit (*SSU*) rRNA gene similarity, are used as theory-agnostic approximations of microbial taxa. Providing impartial partitions of complex sequence datasets, OTUs are the backbone of established workflows for the ecological characterization of microbial communities, such as *mothur* [15] or *QIIME* [16]. Several methods have been developed for binning SSU sequences, most prominently *hierarchical clustering algorithms* (*HCA*, implemented e.g. in *mothur*) and their heuristic approximations, such as *uclust* [17], *cd-hit* [18] or the *ESPRIT* suite of algorithms [19,20]. However, it has been noted that different clustering methods often provide highly inequivalent partitions of the same data, both quantitatively (with respect to total cluster counts and OTU size distributions) and qualitatively (with respect to cluster composition) [21–24]. Consequently, several studies have evaluated approaches to SSU clustering, focusing on distinct measures of cluster quality. Probably the most straightforward test for OTU partition quality has been the comparison of total OTU counts between methods, based on simulated or experimental samples of known composition [19,21,24,25]. Schloss & Westcott [22] used Matthew's Correlation Coefficient as an internal measure of partition quality, based on cluster composition. Alternatively, methods have been

Author Summary

To characterize the composition of microbial communities, researchers often sequence and quantify specific marker genes, particularly the SSU ('small subunit') ribosomal RNA gene. One crucial step in such studies is the clustering of sequences into *Operational Taxonomic Units* (OTUs) of closely related organisms. However, this practice has repeatedly been called into question, arguing that the use of OTUs is not backed by microbial speciation theory. Here, we explore whether OTUs group ecologically similar organisms and show that indeed, OTUs are generally ecologically consistent. Moreover, we show how ecological consistency can be used as a measure of OTU 'quality' and compare different widely used OTU clustering methods. Our findings should help in the design and interpretation of SSU-based microbial ecology studies, in a research field that is only beginning to unfold its full potential to help understand life at the smallest scales.

benchmarked against taxonomically typed ground truth partitions, using measures such as *Variation of Information* (*VI*, [26]), *Normalized Mutual Information* (*NMI*, [20,23,24]) or cluster *Purity* [24] to assess taxonomic consistency. This optimization for taxonomically 'pure' clusters is attractive under the assumption that taxonomic consistency implies both phylogenetic and ecological consistency. However, existing taxon delineations may frequently conflict with phylogeny or refer to ecologically heterogeneous groups of organisms [10], and conflicts between available reference taxonomies, as well as database bias, further reduce the indicative power of taxonomic labels when describing broad ranges of microbial diversity. Moreover, it has been shown that both *NMI* and *VI* produce shifting baseline values, depending on the number of clusters investigated [27], an effect that none of the above-mentioned studies corrects for. Finally, relying on simulated or experimental mixes of known composition as defined inputs may run the danger of missing fundamental challenges brought on by real-world samples (such as microheterogeneity, long-tailed abundance distributions, cellular debris, chimeric molecules, contaminations, etc.). Thus, while taxonomic 'ground truth' may often give a reasonable first assessment, what are alternative and more generally applicable parameters for characterizing 'good' basic units of diversity in microbial ecology?

In this study, we explore the ecological consistency of OTUs. We first revisit and confirm the observation that ecological preferences of microbial lineages are deeply rooted in phylogeny: organisms that share a high SSU sequence similarity tend to be ecologically more similar than expected by chance. We then explore whether this signal is captured by SSU-based OTUs: do organisms that cluster into the same OTU share similar ecological affiliations? In other words, are OTUs ecologically consistent? We approach these questions by first providing anecdotal evidence, before then introducing an *Ecological Consistency Score (ECS)* to provide a more thorough evaluation of OTU ecological consistency. Using a global dataset of roughly one million near full-length SSU sequences, we compare different widely used methods for SSU clustering with respect to how ecologically consistent the OTUs are that they generate. Finally, we reflect on the validity and usefulness of SSU-based OTUs as fundamental units of microbial diversity in light of their ecological consistency, and discuss the implications of using ecological consistency as a taxonomy-independent measure of clustering quality.

Methods

Sequence data & preprocessing

To obtain a comprehensive global dataset, we extracted all full-length 16S/18S rRNA sequences from NCBI *GenBank* ([28], accessed in April 2012) and from the genomes available in the NCBI Reference Sequence Database (*RefSeq* [29], accessed in March 2012). After using Infernal to align sequences to reference consensus models of the bacterial, archaeal and eukaryotic 16S/18S rRNA molecules (provided in the package *ssu-align* [30,31]) and after removing ~20% of total reads that were flagged as chimeric by UCHIME [32], we pruned away any terminal nucleotides that aligned outside of two manually chosen, well-conserved start- and end-positions in the alignment. After these steps, our dataset comprised 950,014 aligned, near full-length sequences (see Text S1 for details).

Sequence clustering into Operational Taxonomic Units

We clustered sequences into OTUs using three HCAs (*average, complete and single linkage*) and two heuristic methods (*uclust, cd-hit*). For every method, we clustered to thresholds of 80–99% sequence identity (92–99% for *average linkage*, see Text S1). We generated OTU sets using *cd-hit* ([18], version 4.5.4, Build 2012-08-25) in *cdhit-est* mode (recommended for clustering highly similar sequences) using standard parameters. The *uclust* ([17], http://drive5.com/usearch/, version 6.0.307) series of OTU sets was generated using the *uclust* software with the *cluster_fast* option and standard parameters. Hierarchical *average*, *complete* and *single linkage* clustering were implemented using the recently developed in-house software package *hpc-clust* [33] using the '*onegap*' sequence distance calculator (counting gaps as single mismatches). *Hpc-clust* parallelizes the hierarchical clustering task and has been shown to cluster sequences as fast as, or even faster than heuristic implementations such as *uclust* and *cd-hit* (less than 3 h wall time for the present dataset of roughly one million sequences on a 256 core computer cluster), while still computing the entire pairwise distance matrix, avoiding any heuristic shortcuts.

Contextual data

We extracted different types of ecologically relevant information from *GenBank* and *RefSeq* annotations. First, we assigned sequences to individual *sampling events* that we define here as unique combination of submitting authors, publication title and isolation source; this classified the dataset into 31,519 samples. Next, we filtered free-text annotations down to 7,202 unique, non-trivial ecological *terms* describing the sampling context. Using a manually curated classification scheme, we annotated samples to 53 more broadly defined *habitat types* (e.g., 'skin' or 'soil', see Text S1 for the full list). In a complementary approach, we filtered annotation keywords for the controlled vocabulary maintained by the Environmental Ontology Project (*EnvO*, http://environmentontology.org/, release date 2011-24-03) and used the ontology to assign related environmental terms to samples (e.g., 'lake' and 'pond' were both classified as 'water body'). This procedure yielded 672 unique *EnvO terms* represented in the dataset. Finally, for samples that are associated with a eukaryotic host, we assigned *host taxonomy* from direct annotations and by inference from annotation keywords. This procedure yielded 2,422 unique host taxonomies (in total representing 5,850 unique taxa) represented in the dataset; remaining archaeal and bacterial sequences were considered *non host-associated*.

Assessing global-scale ecological consistency of OTUs

We developed an *Ecological Consistency Score (ECS)* to assess the ecological consistency of entire sets of sequence clusters with respect to different ecological signals (such as *ecological terms*, see above). The *ECS* was calculated as follows. Consider a partition of a SSU sequence dataset into N OTUs of sizes $n_1, n_2, \ldots, n_N$. What is the likelihood that an ecological feature j with a background frequency of p_j in the entire dataset is observed exactly $k_{i,j}$ times in OTU i of size n_i? We calculated this likelihood $L_{i,j}$ using a binomial model:

$$L_{i,j} = \binom{n_i}{k_{i,j}} p_j^{k_{i,j}} (1-p_j)^{n_i - k_{i,j}}$$

For example, observing 5 sequences annotated with the ecological term 'skin' (background frequency of 30.0%) in an OTU containing 15 sequences has a likelihood of 0.206, but observing the much less frequent term 'hydrothermal' (background frequency ~0.9%) exactly 5 times in the same OTU is much less likely ($L_{15,hydrothermal} = 1.6*10^{-7}$). Similarly, *not* observing a frequent term such as 'skin' in the same OTU has a rather low likelihood ($L_{15,skin} = 0.005$). Thus, the presence of 5 sequences annotated as 'hydrothermal' in an OTU of size 15 is an *enrichment of ecologically similar organisms*, while the absence of a frequent term such as 'skin' in the same OTU is a *negative enrichment*. We computed the summed log-likelihood LL_{set} of the entire partition from the enrichment of every term j in every OTU i:

$$LL_{set} = \sum_i \sum_j \log(L_{i,j})$$

High absolute values of LL_{set} indicated that the distribution of ecological features across the various OTUs in the entire partition were non-random. However, the absolute value of LL_{set} is influenced by total OTU count (as the number of summands i) and OTU size distribution (as n_i in the binomial coefficient). We used an empirical approach to control for these effects: we computed the log-likelihoods LL_{rand} of 1,000 randomized sets with identical cluster size distribution and total count, but with shuffled sequence-to-OTU mapping. This generated a (near-Gaussian) background distribution of LL_{rand}, from which we calculated the *ECS* of the observed OTU set as standard Z score:

$$ECS = -\frac{LL_{set} - \mu_{rand}}{\sigma_{rand}}$$

where μ_{rand} is the average value of LL_{rand} and σ_{rand} is the standard deviation. Thus, *ECS* values indicate by how many standard deviations the enrichment of ecological features in the observed OTU set is removed from a randomized background. In other words, the *ECS* indicates how consistent a given set of OTUs is with respect to an ecological signal, such as the distribution of ecological terms.

Results

SSU similarity is indicative of ecological similarity, and vice versa

Several recent studies have shown that microbes can be remarkably *niche conservative*: ecological affiliations such as habitat preferences are rooted deeply in the tree of life [34,35]. As a consequence of this 'ecological coherence of high bacterial taxa', a close relationship between ecological similarity and SSU similarity has been observed. We confirmed this relationship by exploring a novel, global sequence dataset of roughly one million near full-length SSU sequences, for which we automatically inferred sampling habitats based on ecologically relevant annotation keywords. We calculated pairwise similarities in SSU sequences, ecological terms and inferred habitats (as Jaccard index) for 20 sets of 10,000 randomly selected sequences, resulting in a total of $\sim 10^9$ pairwise comparisons; the results are shown in Figure 1A. For both ecological terms and inferred habitats, we observed a clear trend towards higher ecological similarity at higher SSU similarity. This observation is in line with previous studies that reported a very similar pattern of increasing ecological similarity with decreasing distance on SSU-based phylogenetic trees [34,36]. Moreover, it is concordant with general niche conservatism in microbes, given that our dataset represents a diverse and global survey of microbial taxa. In other words, phylogenetic distance is indicative of ecological similarity. But is the reverse also true? Are ecologically coherent groups of organisms more similar in SSU sequence similarity than expected by chance?

To assess the *internal* SSU similarity of ecologically coherent groups of organisms, we reanalyzed the *human skin microbiome* (HSM) dataset that provides ~100,000 near full-length 16S sequences sampled from distinct body sites [37]. Considering each body site as a unique habitat, we calculated pairwise 16S sequence similarities per sample; the results are shown in Figure 1B, Figures S1, S2 and Table S1. All habitats showed a major abundance of sequence pairs in the 70–80% 16S similarity range, likely corresponding to comparisons of organisms from different bacterial phyla. However, several habitats showed distinctly bimodal (e.g. back, toe web space) or multimodal (e.g. nare, manubrium) distributions of internal 16S similarities, indicating an abundance of more closely related organisms (Figure 1B, top panel). Indeed, these observations are in line with the habitat-wise diversity estimates provided in the original HSM study [37]. When compared to a global background dataset of bacterial 16S sequences (Figure 1B, bottom panel), all skin habitats showed both a notable overrepresentation of highly similar sequence pairs (>90% 16S similarity), as well as the complete absence of a 'tail' of highly dissimilar pairs (<60% 16S similarity). In other words, organisms sampled from a defined skin habitat were more similar to each other in 16S sequence than expected for a global background; this enrichment was statistically highly significant ($p << 10^{-16}$, one-sided Mann-Whitney-U test, see Table S1). The same was true for more broadly defined habitat types: 16S sequences sampled from 'moist', 'dry' and 'sebaceous' skin sites (as classified in the original HSM study) shared significantly higher similarity than expected for a background set (Figure 1B, middle panel, Figure S2 and Table S1). This indicates that in spite of local diversity and distinct internal 16S similarity profiles, the different ecologically coherent habitats (body sites, skin habitat types) sustained communities containing more closely related organisms (higher 16S similarity) than expected for a global background.

Taken together, these results confirm a close relationship between ecological and SSU similarity: closely related organisms tend to be ecologically more similar than expected by chance. However, the reverse is also true: ecological similarity is often indicative of increased SSU similarity.

OTUs are ecologically homogenous on a broad ecological scale

How does this relation between ecological and SSU similarity translate to Operational Taxonomic Units? Are clusters defined by

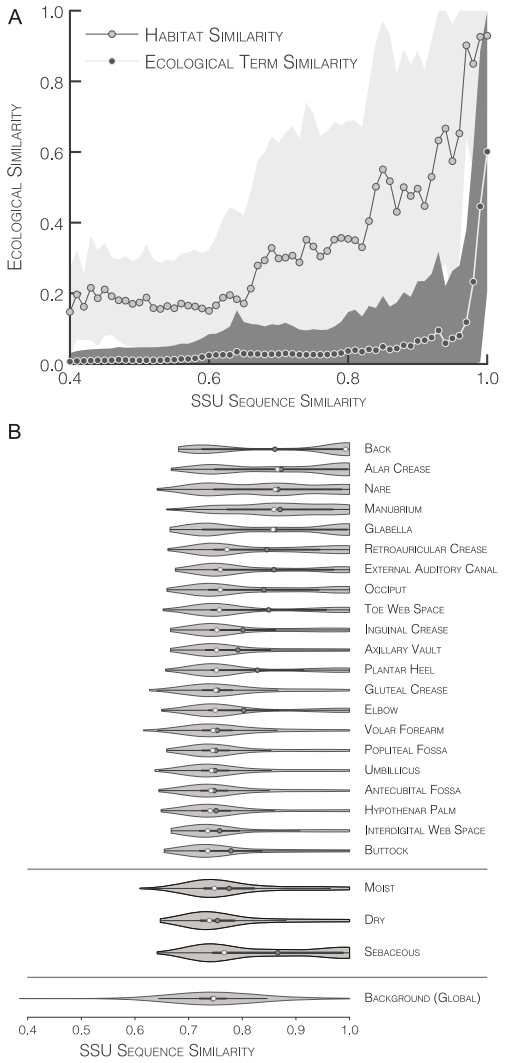


Figure 1. Phylogenetic similarity vs. ecological similarity. (A) General correspondence of ecological and SSU similarity. From our global dataset of roughly one million SSU sequences, 20 datasets of 10,000 sequences each were randomly sampled. For each subset, pairwise sequence similarities and ecological similarities (as *Jaccard Index* of shared annotated ecological terms and habitat types, respectively) were calculated, and the results were averaged over the 20 sets before plotting; mean standard deviations across sets are indicated by grey shades. (B) Internal 16S SSU similarity of human skin habitats. For the *human skin microbiome* dataset [37], pairwise SSU similarities were calculated for all sequences sampled from respective human skin habitats (top) and for sequences from habitats of the same type ('moist', 'dry' or 'sebaceous', as classified by Grice et al [37]; middle). Global background similarities were obtained by calculating pairwise internal SSU similarities for 20 sets of 10,000 sequences randomly drawn from our environmentally heterogeneous set of roughly one million SSU sequences (bottom). Smoothened distributions were drawn based on 150,000 randomly sampled pairwise distances. White circles indicate median, grey circles mean similarity. Non-smoothened, detailed distributions are available in Figures S1 and S2.
doi:10.1371/journal.pcbi.1003594.g001

SSU similarity ecologically consistent? To approach these questions, we clustered a global dataset of roughly one million SSU sequences into OTUs according to different methods that implement fundamentally different clustering regimes. *Hierarchical Clustering Algorithms (HCAs)* compute an entire matrix of pairwise sequence distances and progressively merge the most similar clusters, while *heuristic* methods provide computationally efficient shortcuts. The *complete linkage* (*cl*, *furthest neighbor*) HCA implements an *exclusive* clustering regime, joining two clusters only if every pairwise similarity between the members of each cluster is above the clustering threshold. In contrast, *single linkage* (*sl*, *nearest neighbor*) is *inclusive*, as clusters are joined as soon as any two of their members share above-threshold similarity. Average linkage (*al*, *average neighbor* or *unweighted pair group method with arithmetic mean*, *UPGMA*) conceptually provides a middle ground between the two, requiring that the average pairwise similarity between all members of two clusters be above the threshold for joining them. The most widely employed *heuristic* methods for SSU sequence clustering are arguably *uclust* [17] and *cd-hit* [18]. *Uclust* defines cluster seed sequences, usually depending on sequence length or abundance in the dataset, to which sequences are subsequently compared and linked if the similarity (computed as number of shared short 'words', or *k-mers* between the sequences) is above the required threshold; note that in consequence, *uclust* combines the three steps of sequence alignment, alignment distance calculation and clustering into one. Similarly, *cd-hit* assigns sequences to representative cluster seeds, but uses a different word-matching algorithm and replaces (even implicit) sequence alignment altogether by the use of indexing tables.

Figure 2A shows the ecological associations of the ten largest OTUs for every method when clustering to 97% SSU sequence similarity. We observed that for all methods except *sl*, the majority of OTUs was ecologically homogenous. Clearly, the dominating habitat in the overall dataset, skin (30% of total sequences), also dominated most of the ten largest OTUs for every method, with gastric and intestinal habitats as the second most important fraction. In particular for *cl* and *uclust*, all studied OTUs except '*uclust* OTU 7' consisted of ≥95% sequences sampled from skin, and almost all remaining sequences in these OTUs were annotated as gastric or intestinal. Similarly, most of the observed *al* and *cd-hit* OTUs were dominated by these habitats, albeit to lower extent and with notable exceptions (*al* OTUs 4 & 7, *cd-hit* OTU 5). In contrast, *sl* produced several large clusters that were ecologically heterogenous (OTUs 4, 7–10), with the dominant habitat representing as little as 26.6% of sequences in *sl* OTU 10.

Figure 2B provides a closer look at *sl* OTU 4. It consisted of 17,462 habitat-typed sequences of highly diverse ecological affiliation; for example, sequences sampled from insect hosts, plant hosts, aquatic environments or soil each accounted for 4–5% of diversity within this OTU. We observed that all other tested methods generated significantly more OTUs from the same 17,462 sequences when clustering in the context of the full global set of roughly one million sequences. Indeed, the observed differences in total OTU counts were in the range of 2–3 orders of magnitude, with *uclust* providing 2,102 OTUs where *sl* provided only one. At the same time, we observed that both *cl* and *uclust* provided ecologically more homogenous partitions of the same sequence set, notably by distributing sequences associated to skin and to gastric/intestinal habitats largely into distinct OTUs. Likewise, *al* and *cd-hit* provided ecologically more consistent OTUs

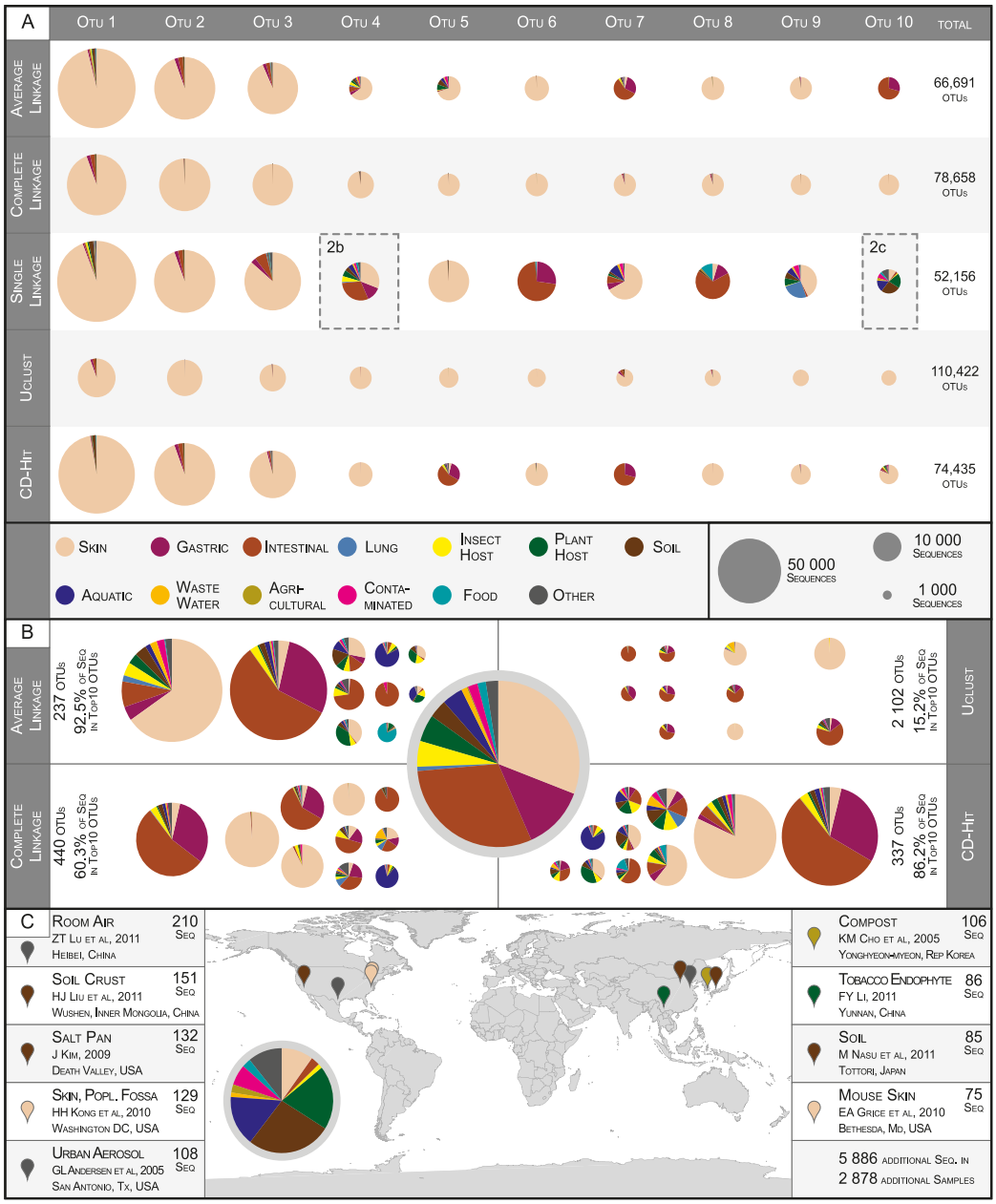


Figure 2. Broad-scale ecological homogeneity of OTUs. (A) Habitat associations of the ten largest OTUs when clustering a comprehensive set of publicly available full-length SSU sequences to 97% similarity using different methods. Pie chart area is proportional to OTU size, colors correspond to habitat types. Total OTU counts are indicated on the right. 9.7% of publicly available sequences lacked habitat annotation, or were typed to conflicting habitats, and were excluded from the analysis. Note that the OTUs shown are not generally identical across clustering methods, but overlap in sequence composition. (B) Breaking down the ecologically inconsistent cluster '*sl* OTU 4'. In the presence of the full global dataset, different methods cluster the 17,462 sequences in *sl* OTU 4 differently, mostly providing ecologically more homogeneous clusters. For every method,

the ten largest clusters and the fraction of sequences they contain, as well as total OTU counts are shown. (C) Sampling events contributing to '*sl* OTU 10'. Geographic locations and isolation sources are shown for nine of the largest sampling events. Marker colors indicate habitat type.
doi:10.1371/journal.pcbi.1003594.g002

than *sl*, albeit to lesser extent. Although all four methods also generated several ecologically heterogenous OTUs, their overall partitions appeared ecologically more homogenous than the single ecologically inconsistent cluster generated by *sl*.

As another example, consider the largest sampling events contributing to *sl* OTU 10 (Figure 2C). Clearly, this OTU contained sequences from very distinct and unrelated ecological contexts, not only on the level of broad habitat types (skin, soil, etc.), but also at finer ecological resolution (e.g., different soil types). Interestingly, this ecological heterogeneity corresponded to a large internal SSU dissimilarity of this particular OTU: although clustered to a nominal similarity threshold of 97%, we observed that a large majority of pairwise similarities within *sl* OTU 10 were actually below this threshold (as can be expected for an *inclusive* clustering algorithm), at a mean internal similarity of 95.2% and with individual pairs of sequences sharing as little as 86% SSU similarity.

The above observations are mostly anecdotal: we considered only a small selection of OTUs and elaborated on individual examples. Nevertheless, this may help to illustrate two important points that will be discussed more rigorously in the following sections: (i) the tested methods clustered the same sequence dataset very differently with respect to total OTU count, OTU size distribution and OTU ecological homogeneity; (ii) with the exception of *sl*, clusters were generally homogenous on a broad ecological scale, considering e.g. that skin and gastric/intestinal habitats are arguably more similar to each other than they are to aquatic or soil habitats.

Global-scale ecological consistency of OTUs depends on clustering method

To refine our above observations on general OTU ecological homogeneity, we developed an *Ecological Consistency Score* (*ECS*, see Methods). Adopting a global perspective rather than focusing on individual examples, the *ECS* is a measure of ecological consistency of entire OTU partitions, taking into account all the clusters provided by a given clustering method. Moreover, focusing on more fine-scale ecological associations than provided by the broadly defined habitat types discussed above, the *ECS* provides increased ecological resolution. High *ECS* values indicate that ecologically similar organisms are clustered, more so than expected by chance.

We tested cluster consistency with respect to four distinct ecological signals: (i) 7,202 *ecological terms* (Figure 3A–C), which we filtered from sequence annotations, provided detailed descriptions of sampling context; (ii) 672 *EnvO terms* (Figure 3D), which we filtered from annotation keywords using the EnvO ontology, provided an alternative and curated hierarchy of ecological descriptions; (iii) *sampling site information* (Figure 3E), for which we considered whether a given OTU contained many sequences that had been sampled from the same site; and (iv) *host taxonomy* (Figure 3F), assuming that closely related host organisms generally provide more similar environments than more distantly related ones. We processed these signals independently, calculating an *ECS* for a given OTU partition for each ecological signal.

We calculated the *ECS* for OTU sets obtained from clustering our global set of roughly one million sequences to nominal similarity thresholds of 80%–99% (92%–99% for *al*, see Text S1) according to different methods: *al*, *cl*, *sl*, *uclust* and *cd-hit* (Figure 3 and Table S2). For all tested datasets, and over the entire range of tested OTU set sizes, we observed similar trends in ecological consistency (*ECS* from highest to lowest): *cl*, *uclust*, *cd-hit*/*al* and *sl*. Over wide ranges of tested OTU counts, differences between OTU definitions were statistically significant (one-sided t-test on jackknifed estimate of *ECS* variability, $p<<0.01$). Jackknifed *ECS* variability was low and constant for all tested datasets and OTU set sizes (coefficient of variation, $0.06<c_V<0.08$).

We observed different and reproducible trends in *ECS* within clustering methods. With increasing clustering stringency (increasing similarity threshold, increasing number of total clusters), *ECS* values monotonically decreased for *cl*, *uclust* and *al*, and for *cd-hit* in the high-cutoff range. This general decrease in ecological consistency might indicate that the rather broad ecological descriptions aligned better with OTUs at lower nominal similarity thresholds, while more closely defined OTUs (higher cluster counts) were not equally well resolved on an ecological scale. In contrast, we observed the opposite trend (decreasing *ECS* with decreasing stringency) for *sl*, and to a lesser extent sometimes *cd-hit*, at lower clustering thresholds. As *sl* is an *inclusive* algorithm (see above), it tends to cluster sequences that share below-threshold similarity. For example, in the previous section we pointed out '*sl* OTU 10', the 10th largest *sl* OTU when clustering to 97% similarity, which clustered sequences sharing below-threshold similarity (mean internal similarity of 95.2%, most dissimilar sequence pair sharing 86% similarity). Since such lumping behavior aggravates with decreasing clustering stringency, it may explain the observed decrease in ecological consistency.

ECS differences between methods were more pronounced with increasing levels of clustering: while at very high similarity thresholds (≥99%), partitions were similar and sometimes indistinguishable on an *ECS* scale, differences of up to ~5-fold between *cl* and *sl* were observed at lower sequence similarity levels. At the frequently-used similarity threshold of 97%, *ECS* scores of *cl* were between 10% and 20% higher than those of *sl*, depending on the feature tested (Table S2). *Cl* also consistently showed the highest *ECS* values when the set of SSU sequences was restricted to those from completed sequenced genomes only (Figure S3). Distinct ecological signals provided different levels of *ECS* resolution: at higher OTU counts, keyword-based measures were less distinctive on an *ECS* scale (ecological term consistency, Figure 3A, and EnvO term consistency, Figure 3D), while sampling site consistency separated OTU definitions better (Figure 3E). Likewise, the archaeal sequence dataset (Figure 3B) distinguished different OTU definitions better than the larger bacterial (Figure 3A) and smaller eukaryal (3C) datasets. However, the general trend was the same across all tested datasets, and across all indicators of ecological consistency: *complete linkage (cl)* generated ecologically more consistent OTUs than the other methods; *single linkage (sl)* resulted in the lowest *ECS* values in all tests; and the remaining methods fell into an intermediate range, while *uclust* generally provided higher ecological consistency than *cd-hit* and *al* which in turn were mostly indistinguishable from each other.

Discussion

Ecological consistency of OTUs is a matter of perspective

Are SSU-based OTUs ecologically consistent? Our results indicate that they are, to a large extent. We detected high levels of ecological consistency both at broad ecological scale in individual

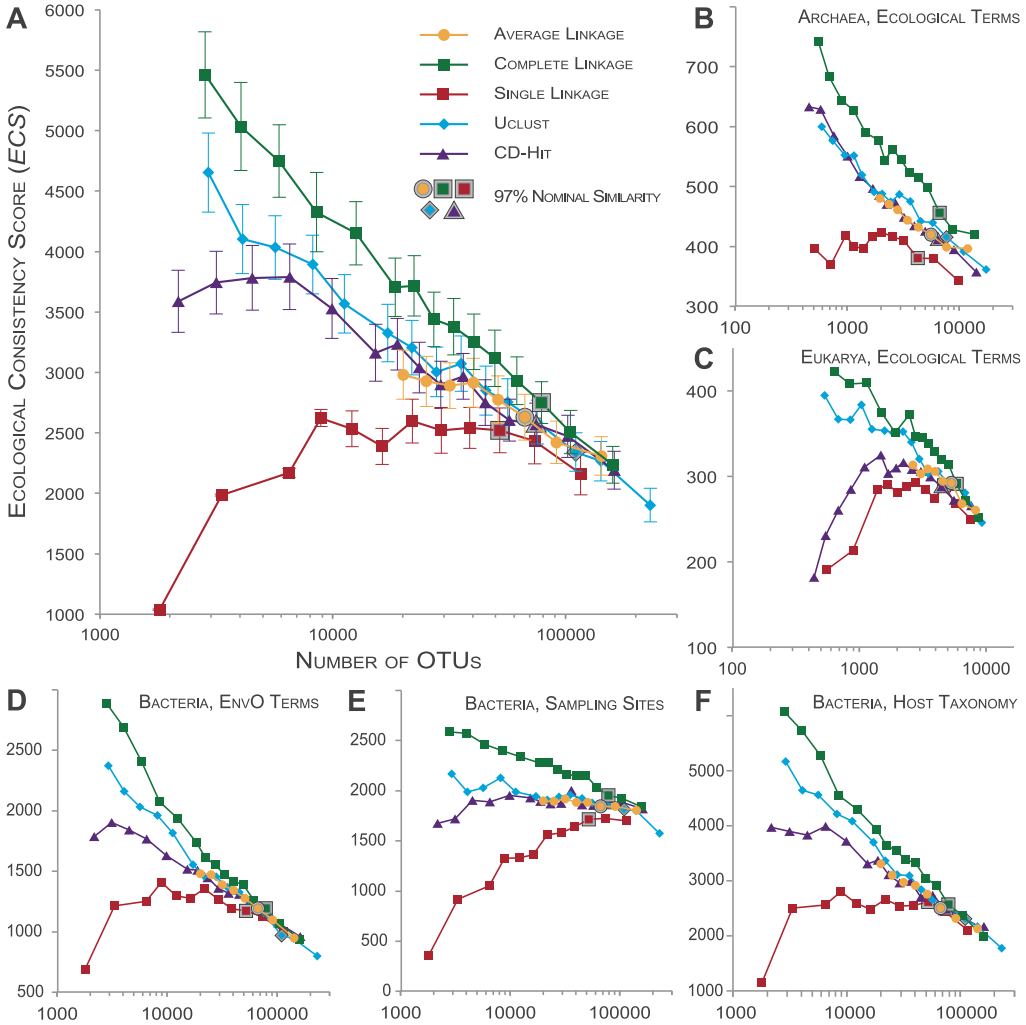


Figure 3. Global *Ecological Consistency Scores* of OTUs. (A) Ecological term consistency when clustering 887,870 *bacterial* full-length 16S sequences according to different methods. *ECS* values (y-axis) describe how non-random the enrichment of ecological affiliations is in a given OTU set (see main text). The total number of clusters including singletons (x-axis) provides for better comparability of methods than nominal clustering thresholds; lower numbers of OTUs correspond to less stringent similarity cutoffs. Error bars indicate jackknifed estimates of *ECS* variability (see Text S1). Data points for OTU sets clustered to 97% nominal sequence similarity are highlighted with a grey shade. The raw data are available in Table S2. For the ecological term consistency when clustering 42,402 archaeal sequences (B), or 20,120 eukaryotic 18S sequences (C), as well as for the bacterial dataset EnvO term consistency (D), sampling site consistency (E), and host taxonomy consistency (F), error bars are not drawn, but variability was in the same range as for (A) (coefficients of variation, $0.06<c_V<0.08$).
doi:10.1371/journal.pcbi.1003594.g003

examples (Figure 2) and at finer ecological scale for a global SSU dataset (Figure 3). In contrast, Koeppel and Wu [13] recently reported an 'extensive ecological heterogeneity among OTUs' for very fine-scale habitat definitions of two model datasets of marine *Vibrio* [5] and hot spring *Synechococcus* [38] communities. Thus, OTU ecological consistency may in fact be a matter of perspective: while OTU clustering may conflict with very high-resolution ecological associations for specific environments, OTUs are generally, though not perfectly, consistent on broader ecological scales. Considering that OTU clustering is a phenomenological approach to diversity analysis, the observed levels of ecological consistency are remarkable: although OTU definitions are mostly independent of underlying assumptions on microbial ecology, they capture groupings of ecologically coherent organisms.

Are the observed levels of ecological consistency sufficient for OTUs to be useful in the ecological characterization of microbial communities? Indeed, it is difficult to globally define appropriate levels of required ecological consistency for 'good' units of microbial diversity. This is largely due to the *ecological plasticity* of microbial taxa at different levels of taxonomic and ecological resolution: while broad-scale ecological coherence in general is deeply rooted in phylogeny [34], several cases of wide 'intra-species' ecological variation have been reported, e.g. within the genera *Bacillus* [39] or *Escherichia* [40]. In other words, though relatedness at family, order or even phylum level is often predictive of a common broad ecological niche, very closely related lineages frequently exhibit surprisingly wide ecological differentiation.

Another frequently cited criterion for biologically meaningful basic diversity units is *phylogenetic consistency*. While Koeppel and Wu recently reported 'extensive and pronounced paraphyly and polyphyly among OTUs' when compared with the *ecotype simulation* algorithm (which uses a phylogenetic tree as input, [13]), we found surprisingly high levels of phylogenetic coherence of *complete linkage* OTUs: with respect to a maximum likelihood tree of 42,024 archaeal sequences, >80% of all non-singleton OTUs at different clustering thresholds were monophyletic (Text S2).

In general, conceptually more sophisticated algorithms to demarcate OTUs such as *ecotype simulation* [4], *CROP* [41] or *M-Pick* [42] may be suited for focused problems, but arguably suffer from throughput problems due to high computational demands (we were not able to execute any of them on our set of one million sequences). On the other hand, impartial OTU clustering conquers large and complex datasets rapidly, while still providing reasonably high levels of ecological consistency. For in-depth studies on broader ecological scopes, OTUs may thus provide good approximations of ecologically coherent lineages.

How good is 'good enough'? Ecological consistency and cluster quality

While we found that OTUs are ecologically consistent in general, there were significant differences between clustering methods. Are these differential levels of ecological consistency indicative of clustering quality? We have shown that an ecological similarity signal, calculated based on contextual data alone, corresponds to SSU similarity for a global, environmentally heterogenous dataset, as well as for the well-defined *human skin microbiome* dataset (Figure 1). Based on this observation, high internal SSU similarity in microbial diversity clusters is expected to correspond to high ecological consistency. In other words, metadata-based ecological consistency can provide a non sequence-based, external measure of cluster quality. Moreover, it is arguably useful to consider ecological consistency when evaluating the quality of diversity units in the context of microbial ecology; nevertheless, ecologically *plastic* diversity units should not be considered inherently 'wrong', since ecological differentiation may occur within groups of closely related organisms. The *Ecological Consistency Score* casts these ideas into an objective framework; it is a global measure of ecological consistency for entire partitions of microbial diversity datasets. Several previous approaches to assessing clustering quality relied on measures such as *Normalized Mutual Information* or *Variation of Information*; these can be problematic, as they are biased by variation in total cluster counts and cluster size distributions [27]. Correcting for these effects, *ECS* values are comparable between different diversity unit definitions.

Considering that our dataset provides a comprehensive survey of microbial diversity, the observed differences in ecological consistency have several interesting implications when interpreted in terms of cluster quality. The tested methods implement different assumptions on the fundamental organization of microbial diversity. Conceptually, *sl* clustering is *inclusive* (guaranteeing that all pairs of above-threshold similarity are clustered, tending to provide fewer and large clusters), while *cl*, *uclust* and *cd-hit* are *exclusive* (preventing any below-threshold pair from clustering and thus tending to provide smaller and more compact clusters); *al*, which focuses on average similarity, provides a balanced middle ground. Our results indicate that exclusive clustering regimes, and in particular *cl*, provide ecologically much more consistent partitions than the inclusive regime of *sl*, and somewhat surprisingly also than *al*. While exclusive and inclusive regimes by definition may provide different partitions at the same nominal similarity threshold in terms of cluster counts, sizes and composition, *ECS* values correct for these effects, in particular when compared across partitions of similar total cluster counts rather than similar nominal sequence similarity. We note that the most rigidly exclusive clustering regime, *uclust*, which at any given threshold provided significantly more (and smaller) OTUs than all other methods, did not provide the highest *ECS* values, probably indicating an over-partitioning of ecologically homogenous clusters.

One potential pitfall of our dataset is sampling bias: clearly, a comprehensive survey of available SSU data will be 'anthropocentric', since in the past, sequencing efforts have been disproportionally concentrated on the human microbiome; for example, 'skin' was the overall most frequent ecological term in the set, annotated to as many as 30% of all sequences. However, the *ECS* framework corrects for potential impacts of this sampling bias by providing the exact same input sequences for each tested method, by using weighted background frequencies for every ecological feature, and by randomizing partitions conservatively. Indeed, our dataset meets many characteristics of reference datasets for *reference-based* approaches to OTU demarcation, as implemented e.g. in QIIME [16]. Such approaches rely on well-defined, comprehensive and usually pre-clustered sets of reference sequences that serve as a 'backbone' to guide the mapping and OTU binning of novel reads. Consequently, the choice of reference pre-clustering method can have a strong impact on resulting reference-based picked OTUs; some of the most commonly used reference sets, provided by the Greengenes [43] and SILVA [44] databases, rely on *uclust* for pre-clustering. As ecological consistency can be an important parameter to optimize for in such globally applicable reference sets, our results may inform the choice of pre-clustering method in such contexts.

Finally, as our findings pertain to global taxonomic and ecological scopes, they are of potential interest for the ongoing debate between taxonomic 'lumpers' and 'splitters' [45–47], considering that exclusive clustering corresponds to 'splitting' regimes, while 'lumping' is inclusive.

When designing a workflow to analyze large sequence datasets, informed choices of methods and parameters are needed at many levels. For example, different denoising protocols, filters for chimeric sequences and alignment methods have previously been benchmarked and are not within the scope of our study. Here, we have focused on sequence clustering into OTUs, and our results may contribute to a more informed choice of clustering method when studying microbial communities: of all tested methods, *complete linkage (cl)* may provide the ecologically most consistent partitions of large sequence datasets. Moreover, there are clearly other aspects of clustering quality that we have not touched upon here, such as robustness to the choice of sequenced SSU gene subregion, portability across studies or the impact of dataset context (does a given method cluster 'rich' and 'sparse' datasets differently?). Nevertheless, ecological consistency is an important

parameter to optimize for, in particular when later using OTUs for the ecological characterization of microbial communities.

To our knowledge, our study provides the first benchmark for SSU clustering methods that employs a signal *external* to both taxonomy and sequence. As more and more environments become available to in-depth ecological characterization, it will be interesting to explore alternative paths towards adopting ecology not only into species concepts, but also into definitions of microbial diversity units. Indeed, our results suggest that 'traditional' OTU clustering has yet an important role to play in this process.

Supporting Information

Figure S1 Pairwise sequence similarities within *human skin microbiome* habitats. This figure contains un-smoothened versions of the sequence similarity distributions shown in Figure 1B. Pairwise internal sequence similarity distributions are shown for every skin habitat from the HSM dataset. Background similarities (indicated in grey) were calculated from 20 sets of 10,000 sequences which were randomly drawn from the global set of bacterial 16S sequences. All similarities were calculated using *hpc-clust* [33].
(PDF)

Figure S2 Sequence similarities within human skin microbiome habitat types. Skin habitats were classified into three types ('moist', 'dry', 'sebaceous') in the original publication by Grice et al [37]. In the upper panel, this figure shows un-smoothened versions of the sequence similarity distributions shown in the middle panel of Figure 1B. Pairwise sequence similarities within habitat types were plotted against similarities between sequences drawn from the global background set (indicated in grey; see Figure S1).
(PDF)

Figure S3 Ecological consistency of OTUs from 4,485 16S gene sequences from fully sequenced genomes. We extracted 4,485 16S genes from fully sequenced genomes downloaded from the RefSeq database [29] and clustered them into OTUs according to different methods (see Methods section in the main text). *ECS* values for all five tested methods are shown; partitions at 97% nominal sequence similarity are highlighted with a grey shade.
(PDF)

Table S1 Sequence similarities within *human skin microbiome* subsets. This table provides the main statistics on sequence similarities for all tested HSM habitats, habitat types and the global background set (Fig. 1B, S1, S2). The rightmost column indicates the p value for a one-sided unpaired Mann-Whitney-U-test of the type '*internal sequence similarity within habitat X is greater than background similarity*'. To calculate internal similarities for the different habitat types (indicated by a star, '*'), 10,000 sequences were randomly drawn from the full sets per habitat type.
(XLSX)

Table S2 Ecological term consistency of clustering methods across similarity thresholds when clustering 887.870 bacterial sequences. 887.870 bacterial sequences were clustered using the hierarchical clustering algorithms *average linkage*, *complete linkage* and *single linkage* (implemented in *hpc-clust*, [33]) and the heuristics *uclust* and *cd-hit*. An Ecological Consistency Score (*ECS*) was calculated with respect to filtered ecological annotation terms. The table reports total OTU counts and ECS values (mean and jack-knifed standard deviation, see Methods in main text); the data corresponds to that shown in Figure 3A in the main text.
(XLSX)

Text S1 Supplementary Methods.
(PDF)

Text S2 Phylogenetic consistency of OTUs. For a global dataset of 42,024 archaeal sequences, *complete linkage* OTUs were tested for monophyly with regard to a maximum likelihood phylogenetic tree.
(PDF)

Acknowledgments

We thank Mark Robinson, University of Zürich, for insightful discussions on the Ecological Consistency Score, as well as Damian Szklarczyk and Alexander Roth for helpful discussions during the preparation of the manuscript. Moreover, we thank three anonymous reviewers for providing highly knowledgeable and deep reviews which greatly helped us improve the present manuscript.

Author Contributions

Conceived and designed the experiments: TSBS CvM. Performed the experiments: TSBS CvM. Analyzed the data: TSBS JFMR. Wrote the paper: TSBS CvM.

References

1. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214. doi:-doi:10.1038/nature11234.
2. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, et al. (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3: 733–739. doi:doi:10.1038/nrmicro1236.
3. Cohan FM (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos T R Soc B 361: 1985–1996. doi:10.1098/rstb.2006.1918.
4. Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, et al. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci USA 105: 2504–2509. doi:10.1073/pnas.0712205105.
5. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, et al. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320: 1081–1085. doi:10.1126/science.1157890.
6. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741–746. doi:10.1126/science.1159388.
7. Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19: 1–7. doi:10.1016/j.tim.2010.10.003.
8. Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ (2013) Distribution-Based Clustering: Using Ecology to Refine the Operational Taxonomic Unit. Appl Environ Microbiol 79: 6593–6603. doi:10.1128/AEM.00342-13.
9. Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biol 7: 116. doi:10.1186/gb-2006-7-9-116.
10. Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19: 744–756. doi:10.1101/gr.086645.108.
11. Doolittle WF (2012) Population Genomics: How Bacterial Species Form and Why They Don't Exist. Current Biology 22: R451–R453. doi:10.1016/j.cub.2012.04.034.
12. Cohan FM, Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. Current Biology 17: R373–R386. doi:10.1016/j.cub.2007.03.032.
13. Koeppel AF, Wu M (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Research 41: 5175–5188. doi:10.1093/nar/gkt241.
14. Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6: 431–440. doi:10.1038/nrmicro1872.
15. Schloss PD, Westcott SL, Rabyn T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75: 7537. doi:10.1128/AEM.01541-09.
16. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7: 335–336. doi:10.1038/nmeth.f.303.
17. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26: 2460–2461.

18. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England) 28: 3150–3152. doi:10.1093/bioinformatics/bts565.
19. Sun Y, Cai Y, Liu L, Yu F, Farrell ML, et al. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Research 37: e76–e76. doi:10.1093/nar/gkp285.
20. Cai Y, Sun Y (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Research 39: e95. doi:10.1093/nar/gkr349.
21. Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12: 1889–1898. doi:10.1111/j.1462-2920.2010.02193.x.
22. Schloss PD, Westcott SL (2011) Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. Appl Environ Microbiol 77: 3219–3226. doi:10.1128/AEM.02810-10.
23. Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, et al. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in bioinformatics 13: 107–121. doi:10.1093/bib/bbr009.
24. Bonder MJ, Abeln S, Zaura E, Brandt BW (2012) Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics (Oxford, England) 28: 2891–2897. doi:10.1093/bioinformatics/bts552.
25. Schloss PD (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLoS Comput Biol 6: e1000844. doi:10.1371/journal.-pcbi.1000844.
26. White JR, Navlakha S, Nagarajan N, Ghodsi M-R, Kingsford C, et al. (2010) Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. BMC Bioinformatics 11: 152. doi:10.1186/1471-2105-11-152.
27. Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison New York, NY: ACM Press. pp. 1073–1080. doi:10.1145/1553374.1553511.
28. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Research 41: D36–D42. doi:10.1093/nar/gks1195.
29. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40: D130–D135. doi:10.1093/nar/gkr1079.
30. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics (Oxford, England) 25: 1335–1337. doi:10.1093/bioinformatics/btp157.
31. Nawrocki EP (2009) Structural RNA Homology Search and Alignment Using Covariance Models. Saint Louis (Missouri): Washington University in Saint Louis, School of Medicine. Available: http://openscholarship.wustl.edu/etd/256/.
32. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics (Oxford, England) 27: 2194–2200. doi:10.1093/bioinformatics/btr381.
33. Matias Rodrigues JF, Mering von C (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. Bioinformatics (Oxford, England) 30: 287–288. doi:10.1093/bioinformatics/btt657.
34. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, et al. (2010) The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol 8: 523–529. doi:10.1038/nrmicro2367.
35. Koeppel AF, Wu M (2012) Lineage-dependent ecological coherence in bacteria. FEMS Microbiol Ecol 81: 574–582. doi:10.1111/j.1574-6941.2012.01387.x.
36. Mering von C, Hugenholtz P, Raes J, Tringe SG, Doerks T, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315: 1126–1130. doi:10.1126/science.1133420.
37. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, et al. (2009) Topographical and Temporal Diversity of the Human Skin Microbiome. Science 324: 1190–1192. doi:10.1126/science.1171700.
38. Becraft ED, Cohan FM, Kühl M, Jensen SI, Ward DM (2011) Fine-scale distribution patterns of Synechococcus ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. Appl Environ Microbiol 77: 7689–7697. doi:10.1128/AEM.05927-11.
39. Maughan H, Van der Auwera G (2011) Bacillus taxonomy in the genomic era finds phenotypes to be essential though often misleading. Infect Genet Evol 11: 789–797. doi:10.1016/j.meegid.2011.02.001.
40. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, et al. (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proceedings of the National Academy of Sciences 108: 7200–7205. doi:10.1073/pnas.1015622108.
41. Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics (Oxford, England) 27: 611–618. doi:10.1093/bioinformatics/btq725.
42. Wang X, Yao J, Sun Y, Mai V (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. BMC Bioinformatics 14: 43. doi:10.1093/nar/gks227.
43. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72: 5069–5072. doi:10.1128/AEM.03006-05.
44. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596. doi:10.1093/nar/gks1219.
45. Ward DM (1998) A natural species concept for prokaryotes. Current opinion in microbiology 1: 271–277. doi:10.1016/S1369-5274(98)80029-5.
46. Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56: 457–487. doi:10.1146/annurev.micro.56.012302.160634.
47. Rosselló-Móra R (2011) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol 14: 318–334. doi:10.1111/j.1462-2920.2011.02599.x.

Supplementary material to manuscript 7.2

Text S1: Supplementary Methods.

This supplementary text is not reprinted here, as it contains information presented in more detail in section 5 (Methods) of this thesis.

Table S2: Ecological term consistency of clustering methods across similarity thresholds when clustering 887.870 bacterial sequences.

887.870 bacterial sequences were clustered using the hierarchical clustering algorithms average linkage, complete linkage and single linkage (implemented in hpc-clust, [33]) and the heuristics uclust and cd-hit. An Ecological Consistency Score (ECS) was calculated with respect to filtered ecological annotation terms. The table reports total OTU counts and ECS values (mean and jack-knifed standard deviation, see Methods in main text); the data corresponds to that shown in Figure 3A in the main text.

Table S2 is a large data table and is not reprinted here.

It is available online (http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003594#s5).

Text S2: Phylogenetic consistency of OTUs.

Phylogenetic consistency is a frequently cited criterion for biologically meaningful basic units of microbial diversity [1-3]. In this view, 'good' diversity unit definitions, complying with evolutionary theory, should provide clusters which are *monophyletic*, rather than *paraphyletic* or *polyphyletic*.

To assess the phylogenetic consistency of OTUs, we analyzed a global dataset of 42,024 near-full length archaeal 16S sequences (see main text and Text S1). We generated a Maximum Likelihood tree assuming a generalized time-reversible (GTR) model using FastTree2 [4]. To this tree, we mapped OTUs obtained from *complete linkage* clustering to different similarity thresholds. As a measure of phylogenetic consistency, we calculated an OTU *monophyly index* with respect to the reference tree. We considered an individual OTU as '100% monophyletic' if (i) all its members shared a single common ancestor, and (ii) no members of the same monophyletic group clustered with any other OTU. To account for different patterns of paraphyly or polyphyly in individual OTUs, we defined P_{anc} as the most recent common ancestor of all sequences pertaining to that OTU. We then calculated 'local monophyly' of the focal OTU as the ratio of sequences belonging to the focal OTU (N_{OTU}) relative to all sequences descending from the most recent common ancestor P_{anc} (N_{desc}) in an approach similar to Koeppel & Wu [5]. For example, an OTU containing 9 sequences which form a paraphyletic group with one additional sequence clustered to another OTU was considered '90% monophyletic'. The overall monophyly index for an entire OTU set was then calculated as the average of the local monophyly of non-singleton OTUs. Note that singleton OTUs (containing only one sequence) are monophyletic by definition, and were not considered when calculating average monophyly, but could locally break monophyly within larger OTUs.

We observed a monophyly index of around 80% for clustering thresholds ≥84% sequence similarity (see data table). These levels of monophyly are remarkably high, in particular when considering that the reference tree itself is probably a close approximation, rather than perfect representation, of the 'true' phylogeny of the tested dataset. We conclude that *complete linkage* hierarchically clustered OTUs are generally, though not perfectly, phylogenetically consistent.

% Sequence Similarity	80	82	84	86	88	90	92	94	96	98	99
Total number of clusters	589	745	958	1,200	1,525	1,973	2,644	3,677	5,381	9,160	13,685
Non-singleton clusters	327	419	544	671	868	1,067	1,392	1,835	2,548	3,670	4,470
Monophyly Index in %	73.1	74.7	80.4	79.7	80.4	80.6	80.3	81.1	82.1	81.3	80.1

References (to Text S2)

1. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, et al. (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3: 733–739. doi:doi:10.1038/nrmicro1236.
2. Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, et al. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci USA 105: 2504–2509. doi:10.1073/pnas.0712205105.
3. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741–746. doi:10.1126/science.1159388.
4. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5: e9490. doi:10.1371/journal.pone.0009490.s003.
5. Koeppel AF, Wu M (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Research 41: 5175–5188. doi:10.1093/nar/gkt241.

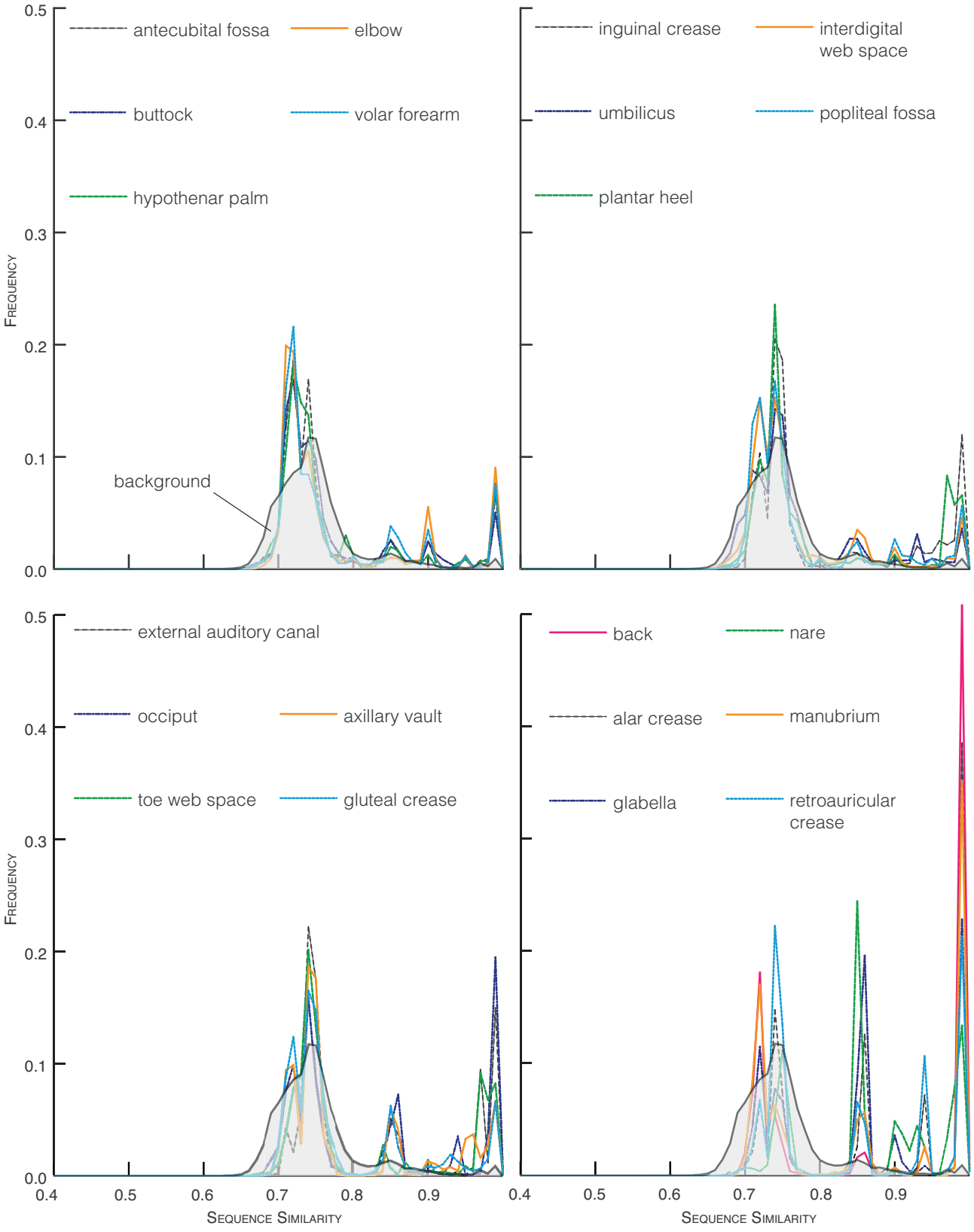


Figure S1: Pairwise sequence similarities within human skin microbiome habitats. This figure contains un-smoothened versions of the sequence similarity distributions shown in Figure 1B. Pairwise internal sequence similarity distributions are shown for every skin habitat from the HSM dataset. Background similarities (indicated in grey) were calculated from 20 sets of 10,000 sequences which were randomly drawn from the global set of bacterial 16S sequences. All similarities were calculated using hpc-clust [33].

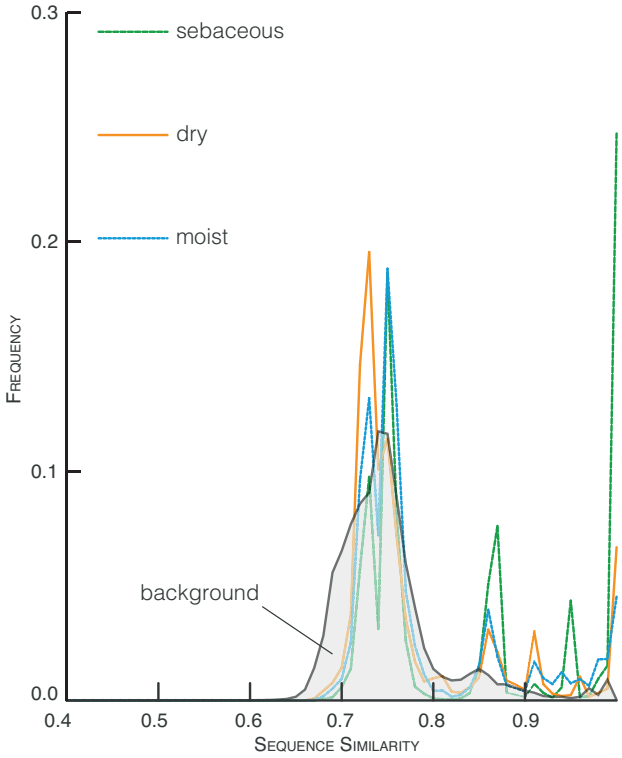


Figure S2: Sequence similarities within human skin microbiome habitat types. Skin habitats were classified into three types ('moist', 'dry', 'sebaceous') in the original publication by Grice et al [37]. In the upper panel, this figure shows un-smoothened versions of the sequence similarity distributions shown in the middle panel of Figure 1B. Pairwise sequence similarities within habitat types were plotted against similarities between sequences drawn from the global background set (indicated in grey; see Figure S1).

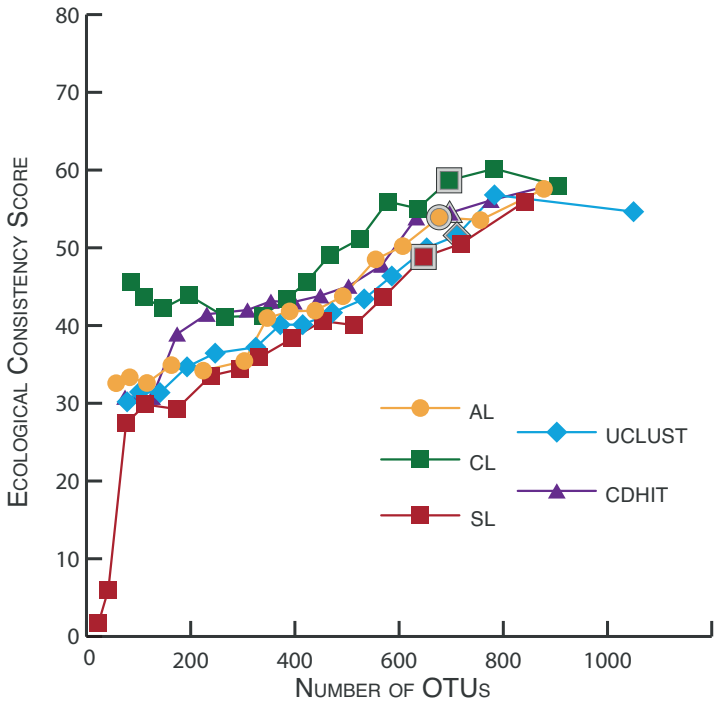


Figure S3: Ecological consistency of OTUs from 4,485 16S gene sequences from fully sequenced genomes. We extracted 4,485 16S genes from fully sequenced genomes downloaded from the RefSeq database [29] and clustered them into OTUs according to different methods (see Methods section in the main text). ECS values for all five tested methods are shown; partitions at 97% nominal sequence similarity are highlighted with a grey shade.

Habitat Name	Number of Sequences	Habitat Type	Mean 16S Similarity	SD	p value *vs global bg*
background (global)	887870	%	0.7554	0.0575	%
dry (pooled)	19634*	dry	0.7755	0.0853	$<< 10^{-16}$
moist (pooled)	42761*	moist	0.7846	0.0854	$<< 10^{-16}$
sebaceous (pooled)	31140*	sebaceous	0.8395	0.1125	$<< 10^{-16}$
volar forearm	10328	dry	0.7801	0.0901	$<< 10^{-16}$
elbow	1916	dry	0.7806	0.0934	$<< 10^{-16}$
hypothenar palm	3371	dry	0.7750	0.0859	$<< 10^{-16}$
buttock	4019	dry	0.7727	0.0807	$<< 10^{-16}$
antecubital fossa	10134	moist	0.7752	0.0848	$<< 10^{-16}$
popliteal fossa	3279	moist	0.7791	0.0851	$<< 10^{-16}$
interdigital web space	3080	moist	0.7747	0.0785	$<< 10^{-16}$
plantar heel	3413	moist	0.8039	0.1017	$<< 10^{-16}$
umbilicus	3737	moist	0.7783	0.0805	$<< 10^{-16}$
inguinal crease	3008	moist	0.8096	0.1032	$<< 10^{-16}$
gluteal crease	3054	moist	0.7912	0.0872	$<< 10^{-16}$
axillary vault	6804	moist	0.8040	0.0967	$<< 10^{-16}$
toe web space	3229	moist	0.8163	0.1057	$<< 10^{-16}$
occiput	7135	sebaceous	0.8267	0.1108	$<< 10^{-16}$
external auditory canal	4399	sebaceous	0.8271	0.1051	$<< 10^{-16}$
retroauricular crease	4208	sebaceous	0.8434	0.1095	$<< 10^{-16}$
manubrium	4035	sebaceous	0.8545	0.1215	$<< 10^{-16}$
nare	3023	moist	0.8748	0.0887	$<< 10^{-16}$
glabella	3496	sebaceous	0.8520	0.1038	$<< 10^{-16}$
alar crease	3417	sebaceous	0.8807	0.1131	$<< 10^{-16}$
back	4450	sebaceous	0.8845	0.1282	$<< 10^{-16}$

Table S1: Sequence similarities within human skin microbiome subsets. This table provides the main statistics on sequence similarities for all tested HSM habitats, habitat types and the global background set (Fig 1B, S1-S2). The rightmost column indicates the p value for a one-sided unpaired Mann-Whitney-U-test of the type 'internal sequence similarity within habitat X is greater than background similarity'. To calculate internal similarities for the different habitat types (indicated by a star, ''), 10,000 sequences were randomly drawn from the full sets per habitat type.*

8. Appendix

8.1 Reprinted co-authored manuscript: Microbiota-Derived Hydrogen Fuels *Salmonella typhimurium* Invasion of the Gut Ecosystem

Maier, L., Vyas, R., Cordova, C. D., Lindsay, H., Schmidt, T. S. B., Brugiroux, S., Periaswamy, B., Bauer, R., Sturm, A., Schreiber, F., von Mering, C., Robinson, M. D., Stecher, B. & Hardt, W. D. (2013)

Cell Host & Microbe, 14(6), 641–651. doi:10.1016/j.chom.2013.11.002

Contribution: TSBS analyzed a targeted 16S sequence dataset to confirm the taxonomic composition of the tested mice' gut microbiota.

This manuscript is available in the printed version of this thesis, and will be available online as soon as permission by the journal publisher is obtained. It is also available directly through the publisher's website:

http://www.sciencedirect.com/science/article/pii/S1931312813004034

Cell Host & Microbe
Article

Microbiota-Derived Hydrogen Fuels *Salmonella* Typhimurium Invasion of the Gut Ecosystem

Lisa Maier,[1] Rounak Vyas,[3] Carmen Dolores Cordova,[1] Helen Lindsay,[3] Thomas Sebastian Benedikt Schmidt,[3] Sandrine Brugiroux,[2] Balamurugan Periaswamy,[1] Rebekka Bauer,[1] Alexander Sturm,[1] Frank Schreiber,[4] Christian von Mering,[3] Mark D. Robinson,[3] Bärbel Stecher,[2] and Wolf-Dietrich Hardt[1,*]
[1]Institute of Microbiology, ETH Zürich, CH-8093 Zurich, Switzerland
[2]Max-von-Pettenkofer Institute, Ludwig-Maximilians-Universität Munich, 80336 Munich, Germany
[3]SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland
[4]Department of Environmental Microbiology, Eawag and Department of Environmental Systems Sciences, ETH Zurich, CH-8600 Dübendorf, Switzerland
*Correspondence: wolf-dietrich.hardt@micro.biol.ethz.ch
http://dx.doi.org/10.1016/j.chom.2013.11.002

SUMMARY

The intestinal microbiota features intricate metabolic interactions involving the breakdown and reuse of host- and diet-derived nutrients. The competition for these resources can limit pathogen growth. Nevertheless, some enteropathogenic bacteria can invade this niche through mechanisms that remain largely unclear. Using a mouse model for *Salmonella* diarrhea and a transposon mutant screen, we discovered that initial growth of *Salmonella* Typhimurium (*S*. Tm) in the unperturbed gut is powered by *S*. Tm *hyb* hydrogenase, which facilitates consumption of hydrogen (H_2), a central intermediate of microbiota metabolism. In competitive infection experiments, a *hyb* mutant exhibited reduced growth early in infection compared to wild-type *S*. Tm, but these differences were lost upon antibiotic-mediated disruption of the host microbiota. Additionally, introducing H_2-consuming bacteria into the microbiota interfered with *hyb*-dependent *S*. Tm growth. Thus, H_2 is an Achilles' heel of microbiota metabolism that can be subverted by pathogens and might offer opportunities to prevent infection.

INTRODUCTION

The mammalian intestine is densely colonized by microorganisms, collectively referred to as microbiota (Ley et al., 2008). The microbiota feature a network of metabolic activities facilitating efficient breakdown of complex diet- and host-derived carbohydrates to short-chain fatty acids (SCFAs), hydrogen (H_2), and carbon dioxide (Fischbach and Sonnenburg, 2011; Flint et al., 2008). Microbial fermentation products are subsequently consumed by crossfeeding secondary fermenters, absorbed by the host, or released into the environment. Gut ecosystem invasion is defined herein as the initial growth phase of a pathogen (or any other newcomer) in the host's intestine. At this stage, the intestinal mucosa appears healthy, and the microbiota is (still) intact and limits nutrient availability. This prohibits growth of most newly arriving bacteria. Despite the scarce nutrient availability, enteropathogens can invade the gut ecosystem. Yet, the factors enabling "gut ecosystem invasion" by enteropathogens remain unclear.

The human food-borne pathogen *Salmonella* Typhimurium (*S*. Tm), a causative agent of diarrhea, can grow up in this nutrient-depleted environment to high numbers and cause disease. Animal experiments established that gut luminal pathogen densities must rise to 10^7–10^8 cfu per gram of stool before enteropathy is elicited (Ackermann et al., 2008; Barthel et al., 2003). As inoculum sizes as low as 10^3–10^5 bacteria suffice for causing diarrheal disease in humans (Food and Agriculture Organization of the United Nations, 2002), we speculated that *S*. Tm can grow initially in the face of an intact microbiota and a healthy gut. The mechanisms fostering *S*. Tm growth in this densely colonized niche are still enigmatic. Such mechanisms can be studied using "low complex microbiota" (LCM) mice, which are permissive for gut luminal *S*. Tm growth (Figure S1A available online; Stecher et al., 2010). LCM mice are ex-germ-free mice which had originally been colonized with strains of the "Altered Schaedler Flora" (Experimental Procedures, Figures S1A and S1E) and permit gut luminal colonization by inocclum sizes as small as 200 colony-forming units (Endt et al., 2010; Stecher et al., 2010). During the first 2 days, there are no signs of enteropathy, and the pathogen grows up to 10^6–10^8 cfu/g stool (gut ecosystem invasion). Mucosal inflammation is elicited at days 3–4 postinfection when the pathogen reaches a final density of 10^8–10^{10} cfu/g stool (Stecher and Hardt, 2011; Figure S1A). Thus, LCM mice should provide a unique model for analyzing all phases of host gut colonization, including gut ecosystem invasion.

RESULTS

Screening for *S*. Tm Mutants Impaired in Early Gut Ecosystem Invasion

To identify *S*. Tm genes required for any stage of gut luminal colonization, we performed an unbiased competitive infection experiment. Specifically, we constructed a set of 500 *S*. Tm transposon mutants (Badarinarayana et al., 2001) and infected LCM mice via the orogastric route. The input pools were

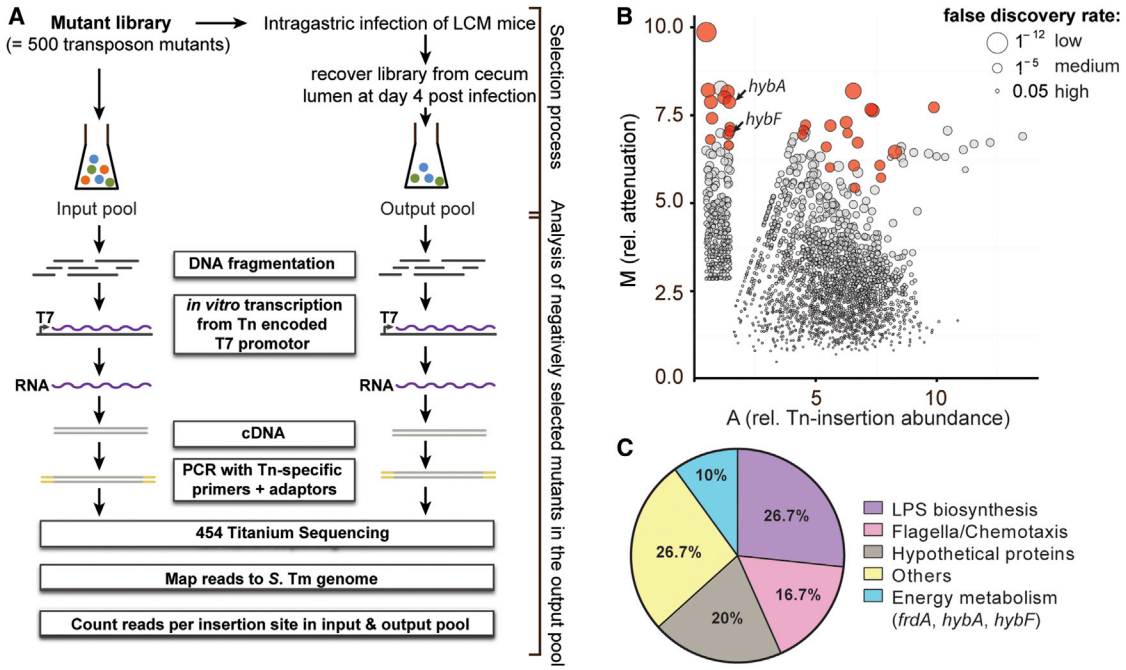


Figure 1. Signature-Tagged Mutagenesis-like Screen for *S.* Tm Genes Required for Gut Lumen Colonization In Vivo

(A) Experimental strategy: 500 randomly generated transposon (Tn) mutants were pooled, and six LCM mice were infected by gavage (Experimental Procedures; Figures S1B–S1E). At day 4 p.i. mutant pools were isolated from the cecum lumen. Next-generation sequencing of transposon-flanking regions using the Tn-encoded T7 promoter permitted identification of Tn insertion sites and of Tn insertions affecting pathogen fitness in the gut lumen.

(B) Statistical analysis of the mutant phenotypes. M/A plot showing the relative attenuation (log_2 fold change in read counts between input and output pools) for each Tn mutant plotted against the relative Tn insertion abundance (= average log_2 counts per million reads, multiplied by the normalized library size to account for differences in the total number of reads sequenced in each sample). A large dot size represents a low false discovery rate (FDR). The 30 most attenuated mutants containing the Tn insertion within a gene are highlighted in red (Table S1). This cutoff was reasonable, as several genes tested in earlier experiments with a C.I. of $0.8 < x < 1.2$ displayed FDR values of $0.005–1^{-5}$.

(C) Functional classification of the 30 most-attenuated Tn insertion mutants.

See also Figure S1 and Table S1.

compared to mutant pools in the cecum lumen at day 4 after infection using transposon-directed insertion-site sequencing (TraDIS; Chaudhuri et al., 2013; van Opijnen and Camilli, 2013), and mutations compromising gut-luminal colonization were identified (six independent animals, two experiments; Figures 1A and S1B–S1E). Transposon insertions in 30 genes reduced gut-luminal abundance of the mutant in all six mice and scored with high confidence ($p \leq 1.3 \times 10^{-5}$; highlighted in red in Figure 1B; Table S1). Almost half of these identified genes were involved in chemotaxis or in flagellar or LPS biosynthesis (Figure 1C). These are well-established *S.* Tm virulence factors required for growth and survival in the inflamed gut (Allen-Vercoe and Woodward, 1999; Chaudhuri et al., 2013; Craven, 1994; Ilg et al., 2009; Stecher et al., 2008; Stecher et al., 2004). These genes likely contribute to expansion/maintenance of the pathogen population at days 3 and 4 of the experiment and confirmed the robustness of our experimental approach. We also identified three genes involved in anaerobic energy metabolism (Figure 1C), *frdA*, the first gene of the operon encoding the fumarate reductase complex, *hybA* and *hybF*. The latter two genes encode subunits of a NiFe-hydrogenase known to consume molecular hydrogen as an electron source in anaerobic environments, thus powering microbial growth ("energy conservation"; Figure S2A) (Lamichhane-Khadka et al., 2010; Maier et al., 2004; Zbell et al., 2008). As H_2 is produced by primary fermenters of the microbiota (*not* the host; Fischbach and Sonnenburg, 2011; Flint et al., 2008), this provided hints that *S.* Tm may capitalize on this microbiota-derived metabolite during some stage of intestinal colonization.

Hydrogen Consumption by *S.* Tm Is Only Required during the Initial Phase of Gut Ecosystem Invasion

In order to verify the role of hydrogenases during gut infection, we constructed site-directed mutants (Figure S2B; Supplemental Experimental Procedures). In competitive infections, the *hyb* mutant (*S.* Tm^{hyb}; *hybBCAhypO*, which lacks all structural genes of the *hyb* hydrogenase) displayed a pronounced growth defect compared to the isogenic wild-type strain (≈100-fold; $p < 0.05$; Figure 2). This was corrobated by *hyb* expression in the gut lumen (Figure S2D). Interestingly, the growth defect of *S.* Tm^{hyb} was restricted to the first day of the experiment when pathogen loads were still low ($\leq 10^8$ cfu/g stool) and no signs of mucosal inflammation were observed (Figures 2B–2D). Thereafter, the competitive index did not drop any further (Figure 2A). These data indicate that *S.* Tm requires *hyb* only in the initial phase of gut ecosystem invasion, but not at later stages of the infection, and that this initial stage (days 0–1) is mechanistically distinct.

Further experiments excluded major contributions of two alternative H_2-consuming hydrogenases encoded in the *S*. Tm genome (Figure S2B; Supplemental Experimental Procedures). Disrupting the two alternative hydrogenases yielded no defects in gut ecosystem invasion, and the hydrogenase triple mutant (*S*. Tm^{hyd3}) displayed the same in vivo growth defect as did *S*. Tm^{hyb} (Figures S3A and S3B). Thus, while *hyb* is necessary for robust pathogen growth in the host's intestine, the other two hydrogenases contribute little. This was further supported by complementation (Figure S3B). Furthermore, the gut ecosystem invasion defect of the hydrogenase mutant was independent of the inoculum size and also observed upon gavage of 5×10^3 cfu (data not shown; standard inoculum size = 5×10^7 cfu; Experimental Procedures). Finally, in vitro experiments in anaerobic broth culture verified that the growth defect of *S*. Tm^{hyd3} was only observed in the presence of H_2, but not in its absence (Figures S4A and S4B). In conclusion, these data confirmed the pivotal importance of *hyb* for H_2-dependent *S*. Tm growth.

Our initial data suggested that the *hyb* hydrogenase may fuel pathogen growth during gut ecosystem invasion, i.e., the first 24 hr p.i. (Figure 2A). At this stage the pathogen grows in the face of the resident microbiota (which presumably still produces H_2) and overt inflammation is not yet triggered (Figures S1A and 2B–2D). To further substantiate the need for hydrogenases in the noninflamed gut, we performed competition experiments in the avirulent strain background. The isogenic *S*. Tm mutant (*S*. Tm^{avir}; Δ*invG*Δ*sseD*; Supplemental Experimental Procedures) colonizes the gut but remains "locked" in gut ecosystem invasion phase of the infection, as it lacks two key virulence factors and therefore cannot elicit overt mucosal inflammation (Hapfelmeier et al., 2005; Stecher et al., 2007). To this end, we constructed a hydrogenase-deficient mutant in the *S*. Tm^{avir} background (*S*. $Tm^{avir\ hyd3}$). First, we tested this strain's capacity to grow up in the gut of LCM mice. In competitive infections, *S*. $Tm^{avir\ hyd3}$ displayed a pronounced colonization defect on day 1 p.i. but no further decrease from day 1 to day 4 p.i. (Figure 3). These results were strikingly similar to those obtained in the wild-type *S*. Tm strain background (compare Figure 2A and Figure 3A) and verified that hydrogenases are indeed only required during gut ecosystem invasion, whether inflammation is triggered or not. Accordingly, intravenous infection experiments confirmed that hydrogenases are not needed for growth at systemic sites (Figure S3C). This further supported the notion that gut ecosystem invasion is a distinct step in host intestinal colonization, which prepares the ground for subsequent stages of the infection.

Microbiota-Derived H_2 Is Responsible for the Competitive Defect of *S*. Tm Hydrogenase Mutants during Early Gut Invasion

Next, we addressed the role of the resident microbiota in *hyb*-dependent gut ecosystem invasion. As the microbiota is considered to be the source of all available H_2, presence of a H_2 producing microbial community should be required for hydrogen-dependent pathogen growth. To this end, we measured H_2 concentrations in freshly dissected ceca ex vivo using a hydrogen microsensor (Experimental Procedures). In germ-free mice lacking all associated microbiota, no H_2 was measurable in the cecum lumen (<2 μM, Figure 4A), and *S*. $Tm^{avir\ hyd3}$ did not display any competitive growth defect (Figures 4B, 4E, and 4F). In contrast, the cecum of LCM mice harbored high levels of H_2 (Figure 4A). This was strikingly similar to the levels of H_2 in the cecum of CON mice, which harbor a "normal," complex microbiota (Figure 4A), as well as the large intestine of humans and diverse animal species (Table S3). With a K_M value of the *S*. Tm hydrogenase activity of 2.1 μM (Maier et al., 2004), these data verified microbiota-derived H_2 as a possible energy source during gut ecosystem invasion in vivo. Indeed, the competitive growth defect of *S*. Tm^{hyd3} in CON mice was comparable to that of LCM mice (Figure 4C and Figure 4D, left side; Figures 4E and 4F). As a complementary approach, we tested the effect of antibiotic pretreatment, a procedure known to reduce microbiota abundance by >10-fold, shift the microbiota composition, and increase metabolite availability in the large intestinal lumen (e.g., carbohydrates like fucose and sialic acids, both accessed by *S*. Tm for intestinal expansion) (Ng et al., 2013; Willing et al., 2011). This should alleviate the need for *hyb*-dependent growth. Indeed, microbiota disruption by streptomycin pretreatment abrogated the competitive growth defect of *S*. Tm^{hyd3} in both LCM and CON mice (Figure 4C and Figure 4D, right side; Figures 4E and 4F). Conversely, microbiota transplantation from LCM mice to another gnotobiotic mouse model (VLCM mice; yield just a small C.I. for *S*. $Tm^{avir\ hyd3}$) reduced the colonization efficiency of *S*. $Tm^{avir\ hyd3}$ in competitive infections (Figures S4C and S4D). Finally, we quantified the total gut luminal population sizes achieved by a hydrogenase-deficient *S*. Tm strain. In both LCM and CON mice, *S*. $Tm^{avir\ hyd3}$ yielded significantly lower total intestinal *Salmonella* loads than the parental strain (*S*. Tm^{avir}; Figure 5). Collectively, these findings support the pivotal role of microbiota-derived H_2 during gut ecosystem invasion by *S*. Tm.

Genes Encoding for H_2-Producing Enzymes Are Abundant in Microbial Gut Metagenomes

Metagenome analyses were performed to assess the potential availability of H_2 in different hosts. Microbial H_2-metabolizing pathways, which are essential for efficient fermentation, are thought to rely on three classes of enzymes: NiFe-hydrogenases, FeFe-hydrogenases, and HmD-like enzymes (Schwartz and Friedrich, 2006). Based on the presence of sequences for one or more of these enzymes, all publically available gut metagenomes showed evidence for H_2-generating pathways (Tables 1 and S4; Experimental Procedures). The same was true for the cecal microbiota of the LCM mice studied here (MG-Rast accession numbers 4535626.3 and 4535627.3). This was well in line with published work on H_2 levels measured in the intestinal tract of animals and man (Table S3) and verified that H_2 production indeed represents a universal metabolic feature of the complex microbiota (and our simplified LCM model). However, the absolute H_2 levels may vary depending on host species or diet. Thus, the balance between H_2 production (i.e., by primary fermenters; Carbonero et al., 2012) and "H_2-loss" by H_2-consuming species of the microbiota (e.g., the methanogens like *Methanobrevibacter smithii*, the reductive acetogens like *Blautia hydrogenotrophica*, and sulfate-reducing bacteria like *Desulfobacter* spp. or *Desulfovibrio* spp.; Carbonero et al., 2012), as well as by diffusion, blood-mediated transport, and

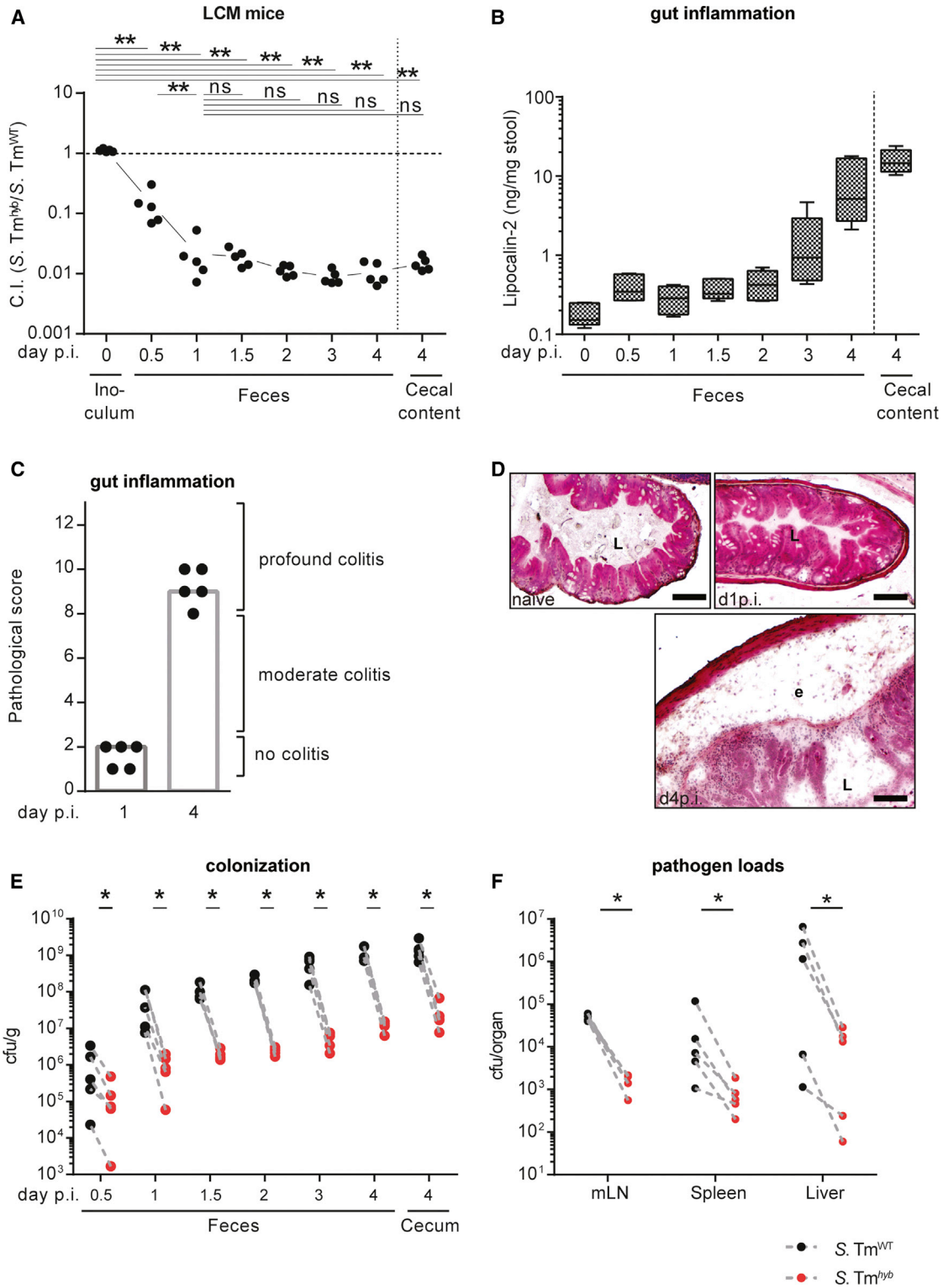


Figure 2. *S.* Tm *hyb*-Hydrogenase Mutant Shows Defective Gut Ecosystem Invasion

(A) Mice were infected with 1:1 mixtures (5×10^7 cfu by gavage) of the *hyb*-hydrogenase mutant and the isogenic hydrogenase-proficient background strain *S.* Tm^{WT}. Fecal loads of both strains were determined by plating and served to calculate of the competitive indices (C.I.s; Experimental Procedures). C.I. experiments were performed in five naive LCM mice. ns, not significant ($p \geq 0.05$), **$p < 0.01$; Mann-Whitney U test.

(B) Lipocalin-2 ELISA monitoring the onset of inflammation during the course of the experiment. Box and whiskers plot: the box indicates first and third quartiles, and whiskers denote minimal and maximal measurement readings.

(legend continued on next page)

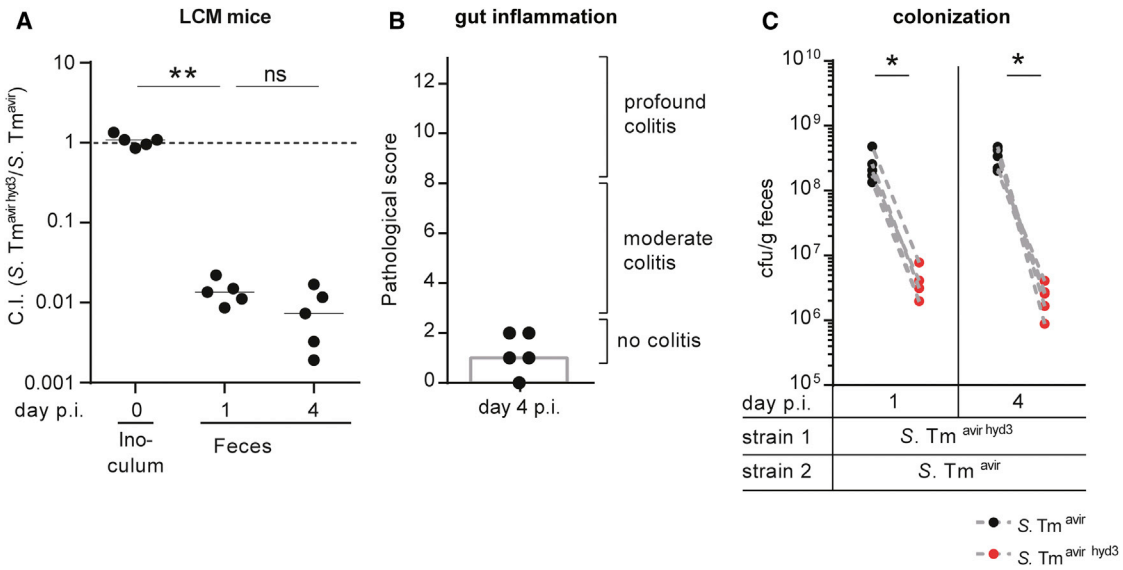


Figure 3. *S.* Tm Only Profits from H_2 during the Initial, Noninflammatory Phase of Gut Ecosystem Invasion

(A) C. I. experiments were performed in five naive LCM mice to test in vivo fitness of *S.* $Tm^{avir\ hyd3}$. ns, not significant ($p \geq 0.05$), **$p < 0.01$; Mann-Whitney U test.

(B) Pathological scores of the cecal mucosa at day 4 p.i. Cecal tissue sections from the competitive infection experiment shown in (A) were stained with HE and scored for inflammation.

(C) Fecal loads of *S.* $Tm^{avir\ hyd3}$ and *S.* Tm^{avir} at day 1 and day 4 p.i. were determined by differential plating. *$p < 0.05$, one-tailed Wilcoxon matched pairs signed rank test on paired data (dashed lines).

See also Figure S3.

exhalation, may dictate the efficiency of gut ecosystem invasion by incoming enteropathogens. As nutrition can affect gut microbiome richness and hydrogen availability (Cotillard et al., 2013; Le Chatelier et al., 2013), infection risks may depend in part on dietary habits.

Addition of an H_2 Consumer Can Interfere with *hyb*-Dependent *S.* Tm Growth

Due to their simplified species composition, the LCM mice offer a unique opportunity to manipulate functional features of the microbiota, e.g., by adding species or shifting the intestinal H_2 balance. To this end, we precolonized LCM mice with an additional "H_2 consumer," *S.* Tm^{avir} (Figure 6A). Control mice were precolonized with *S.* $Tm^{avir\ hyd3}$, a *S.* Tm strain which cannot consume hydrogen. In subsequent competive infection experiments, hydrogenases proved to be of greater importance for gut ecosystem invasion in the control mice than in the mice precolonized with *S.* Tm^{avir} ($p < 0.05$; *S.* $Tm^{avir\ hyd3}$ versus *S.* Tm^{avir}; Figures 6B and S5). Thus, pathogen colonization could be thwarted by introducing a H_2 consumer. This further supported the key role of H_2 for the initiation of *S.* Tm infection.

DISCUSSION

Our findings establish gut ecosystem invasion as a critical step of the orogastric *S.* Tm infection. During this initial phase of the infection, pathogen growth in the gut relies at least in part on metabolites provided by the microbiota. This differs markedly from the interactions observed later (i.e., during expansion/maintenance), when the host's mucosal immune response fuels pathogen growth and suppresses the microbiota (Kaiser et al., 2012; Winter et al., 2013). Thus, colonization of the host's gut comprises different phases featuring distinct sets of positive and negative interactions. The interactions between the pathogen, the microbiota, and the host are clearly more complex than previously anticipated.

Gut ecosystem invasion by *S.* Tm relies on H_2. This is true for mice harboring two different microbiotas of reduced complexity (LCM mice used thoughout most of this study; VLCM mice used in Figures S4C and S4D) or animals with a normal SPF microbiota, alike (Figures 4D–4F and 5B). In contrast, intravenous infections did not yield any evidence for H_2-dependent pathogen growth at systemic sites (Figure S3C). At first sight, this seems

(C and D) Histopathological evaluation of HE-stained cecal sections (L, intestinal lumen; e, edema in submucosa) of these mice. The HE-stained cecal tissue for day 1 p.i. was taken from the experiment shown in Figure S3A (1:1 infection with *S.* Tm^{WT} and *S.* Tm^{hyb}). Scale bar, 100 μm. This demonstrated that mucosal inflammation was elicited at days 3–4 postinfection, as confirmed by pathological scoring.

(E) The bacterial loads of *S.* Tm^{WT} (black symbols) and *S.* Tm^{hyb} (red symbols) populations were monitored in the feces during the course of the infection and in the cecal content at the end of the experiment. These data verify the distinct colonization defect of *S.* Tm^{hyb} during the first day of infection.

(F) Pathogen loads of *S.* Tm^{WT} (black symbols) and *S.* Tm^{hyb} (red symbols) in systemic organs at day 4 p.i. *$p < 0.05$, one-tailed Wilcoxon matched pairs signed rank test on paired data (dashed lines). Please note that the reduced loads of *S.* Tm^{hyb} in lymph nodes, spleens, and livers were most likely attributable to the reduced seeding from the intestinal lumen (which must have occurred after the initial *hyb*-dependent growth in the gut; see Figure S3C, below).

See also Figure S2 and Table S2.

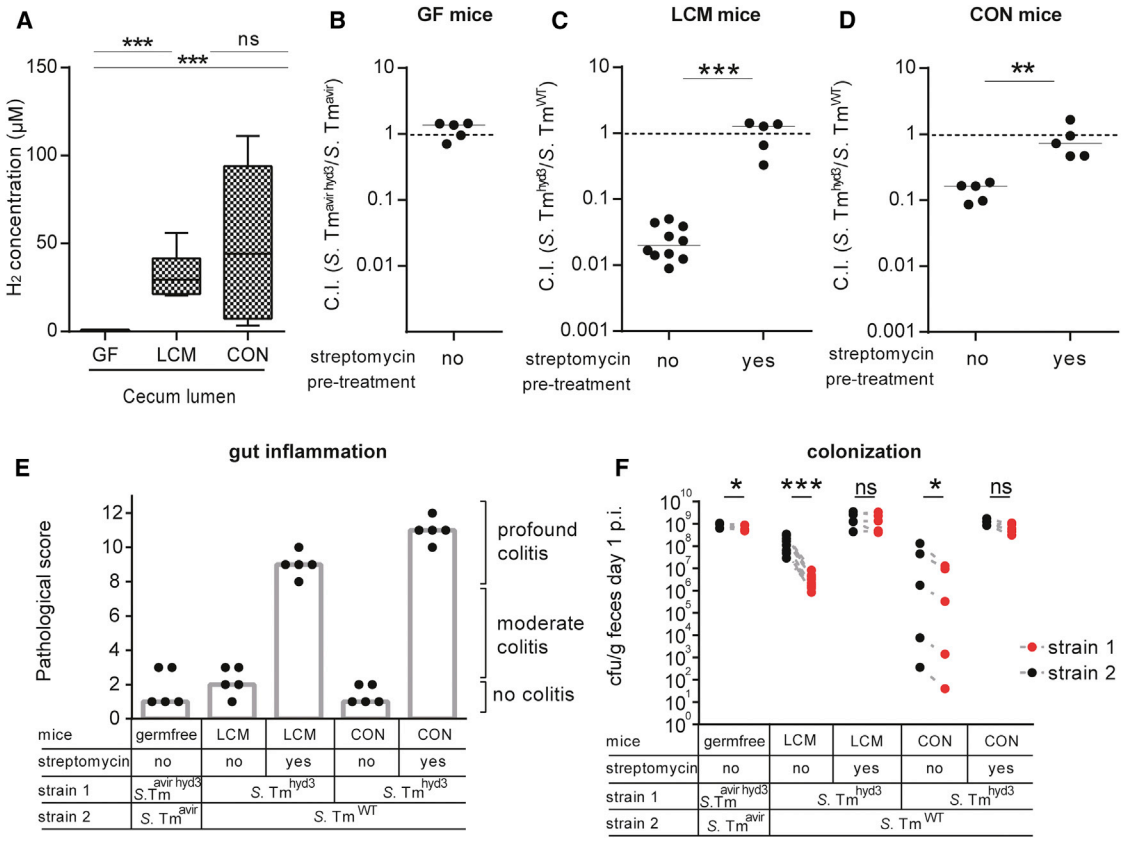


Figure 4. Defective Gut Ecosystem Invasion by *S.* Tm Hydrogenase Mutants Is Linked to Microbiota-Derived H_2

(A) H_2 levels in the cecum lumen. H_2 concentrations were measured at three different positions in the cecum and corrected for electrode crosssensitivity to H_2S (≥3 mice per group) (Experimental Procedures). Box and whiskers plot: the box indicates first and third quartiles, and whiskers denote minimal and maximal measurement readings.

(B) C.I. experiment of *S.* $Tm^{avir\ hyd3}$ versus *S.* Tm^{avir} in five germ-free mice (5×10^7 cfu by gavage; analysis at day 1 p.i.).

(C) C.I. experiment of *S.* Tm^{hyd3} versus *S.* Tm^{WT} in naive LCM mice or streptomycin pretreated animals (10/5 mice per group; 5×10^7 cfu by gavage; analysis at day 1 p.i.).

(D) C.I. experiment of *S.* Tm^{hyd3} versus *S.* Tm^{WT} in naive CON mice or streptomycin pretreated animals (five mice per group; 5×10^7 cfu by gavage; analysis at day 1 p.i.). ns, not significant ($p \geq 0.05$), **$p < 0.01$, ***$p < 0.001$; Mann-Whitney U test.

(E) Pathological scores of the cecal mucosa at day 1 p.i. Cecal tissue sections from the competitive infection experiment shown in (B)–(D) were stained with HE and scored for inflammation.

(F) Bacterial loads of both competing strains at day 1 p.i. were determined by differential plating. ns, not significant ($p > 0.05$), *$p < 0.05$, ***$p < 0.001$; one-tailed Wilcoxon matched pairs signed rank test on paired data (dashed lines).

See also Figure S4 and Table S3.

to be in conflict with earlier work in the oral infection model for typhoid fever (Maier et al., 2004). Upon oral infection, hydrogenase mutants of *S.* Typhimurium ATCC14028 failed to colonize the livers and spleens. Our data may suggest that this attenuation was attributable at least in part to defective growth in the gut, before the bacteria had actually disseminated to systemic sites. This hypothesis would be in line with hydrogenase expression of ATCC14028 in the murine ileum (Zbell et al., 2008). However, we cannot formally exclude that ATCC14028 differs from the SL1344 strain used in our study in being capable of utilizing H_2 in liver and spleen. Such strain-specific differences may affect the adaptation to new hosts. Clearly, *S.* Tm SL1344 requires H_2 only for gut colonization, but not at systemic sites (Figure S3C). This provides a striking example for a central intermediate of microbiota metabolism fuelling pathogen growth at a site occupied by a dense commensal community. Due to the conserved nature of the metabolic network of the gut microbiota, this metabolite will likely be available in any host animal as well as in humans. Thus, H_2 could be regarded as an "Achilles' heel" of microbiota metabolism which can be exploited by *S.* Tm for gut ecosystem invasion.

Molecular hydrogen might affect a number of enteric bacterial infections. This is indicated by genetic evidence for hydrogen-consuming hydrogenases, in vitro data demonstrating roles of hydrogenases in energy conservation, metabolite uptake, and acid resistance by various enteropathogens, including *E. coli*, *Shigella* spp., *Yersinia* spp., and *Campylobacter* spp. (Lamichhane-Khadka et al., 2011; Lamichhane-Khadka et al., 2010; Maier, 2005; Maier et al., 1996; McNorton and Maier, 2012; Zbell et al., 2007; Zbell and Maier, 2009) (Table S2), and by groundbreaking in vivo experimentation on *Helicobacter pylori* (Maier, 2003; Olson and Maier, 2002). The latter requires an

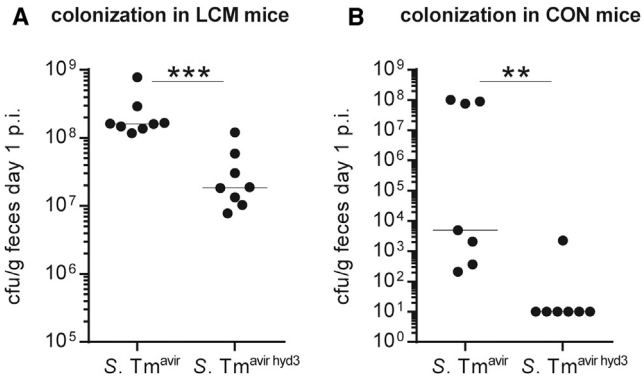


Figure 5. *S.* Tm $^{avir\ hyd3}$ Is Impaired in Colonization of Naive LCM and CON Mice

(A) Eight naive LCM mice or (B) seven naive CON mice were infected with either *S.* Tm^{avir} or *S.* Tm$^{avir\ hyd3}$ (5×10^7 cfu by gavage), and fecal loads were determined at day 1 p.i. $^{**}p < 0.01$, $^{***}p < 0.001$; Mann-Whitney U test.

uptake-type hydrogenase for H_2-dependent colonization of the murine stomach. Interestingly, the H_2 measured at this site was thought (though never shown) to derive from the large-intestinal microbiota. In contrast to the large intestine, which features microbiota densities of 10^{12} cfu/g stool, the stomach is typically colonized by no more than 10^1 microbial cells per gram of content (Sommer and Bäckhed, 2013). Thus, the high diffusibility of H_2 between different organ systems may explain how microbiota-derived H_2 can be tapped not only by pathogens (like *S.* Tm) growing among (and finally outcompeting) the microbiota in the large intestine but also by pathogens colonizing sterile (or almost sterile) sites.

The manipulation of essential metabolite availability may help in preventing pathogen colonization. In fact, as common practice, broiler chicks are treated with attenuated *Salmonella* spp. to reduce the incidence of pathogenic *Salmonella* spp. (Kerr et al., 2013). It is tempting to speculate that this "competitive exclusion" strategy is based at least in part on reduced local availability of H_2. As other enteropathogenic bacteria are also equipped with hydrogenases, H_2 exploitation may represent a common strategy for colonizing the gut. The molecular understanding of the gut ecosystem invasion phase might reveal unique opportunities for thwarting pathogen colonization right from the beginning.

EXPERIMENTAL PROCEDURES

Bacterial Strains

All *S. enterica* serovar Typhimurium strains used in this study are derivatives of the streptomycin-resistant wild-type strain SL1344 (SB300) (Hoiseth and Stocker, 1981) (Supplemental Experimental Procedures). Deletions in the hydrogenase genes were constructed using the lambda/red homologous recombination technique (Datsenko and Wanner, 2000). The genomic region to be deleted was substituted by a *cat* cassette from pKD3 or *aphT* from pKD4. After P22 phage transduction of the antibiotic resistance-substituted region into a clean SB300 strain, the cassette was removed using pCP20 encoded flippase (if indicated). For complementation of the *S.* Tm^{hyb} mutation, the gene SL1344_3112 encoding for a hypothetical protein was substituted by a cat cassette using a lambda/red recombination approach. Substitution of SL1344_3112 with an antibiotic resistance marker did not affect in vivo fitness of the strain (data not shown). P22 phage transduction of the marker including intact *hybABChypO* region into the mutant strain was performed to insert a functional copy of the deleted genomic region into the mutant strain. All constructs were verified by PCR.

Animal Experiments

Animals: CON, LCM, and GF

All animals used in this study are C57BL/6 mice associated with different types of microbiota. Conventional (CON) mice are mice from our in-house colony at the Rodent Center HCI (RCHCI) (Zurich, Switzerland) under specified opportunistic and pathogen-free conditions in individually ventilated cages. LCM (low complex microbiota) mice are ex-germ-free mice which were colonized with the members of the Altered Schaedler flora in 2007 (Stecher et al., 2010) and ever since bred under strict hygienic isolation in a separate breeding room. VLCM (very low complex microbiota) mice are bred at Max-von-Pettenkofer Institute (Munich, Germany) and were generated by inoculating germfree C57BL/6 mice with three strains of the Altered Schaedler flora (ASF361, ASF457, and ASF519; Dewhirst et al., 1999) as pure culture. Germfree C57BL/6 mice were generously provided by the University Hospital Bern. Each experiment was performed at least twice independently, and the data were pooled.

Infection and Competitive Infection Experiments

Single-infection and coinfection experiments were performed in 8- to 12-week-old mice with different composition of the microbiota. Mice were infected as described previously (Barthel et al., 2003). Pretreatment with 20 mg streptomycin was only performed if indicated (i.e., Figures 4C and 4D, right panels; Figures 4E and 4F). For infection or colonization, bacteria were grown for 12 hr in 0.3 M NaCl supplemented LB medium containing the appropriate antibiotic(s), diluted 1:20, and subcultured for 4 hr in the same medium without supplement of antibiotics. Mice were infected with 5×10^7 bacteria by gavage. Freshly collected fecal pellets were harvested, and homogenized in PBS with steel balls in a tissue lyser (QIAGEN) for plating (and frozen for lipocalin-2 ELISA analysis; inflammation marker). Differential plating on MacConkey agar plates (Oxoid) supplemented with the appropriate antibiotics (50 μg/mL streptomycin, 30 μg/mL kanamycin, 30 μg/mL chloramphenicol, 100 μg/mL ampicillin, 12 μg/mL tetracycline) allowed determination of bacterial population size. The competitive index was calculated by dividing the population size of the mutant strain by the population size of the corresponding background strain. The result was corrected for the ratio of both strains in the inoculum. For quantifying live bacterial loads in the organs, mice were sacrificed by cervical dislocation at the indicated time point (untreated, day 1 p.i., day 4 p.i.), and cecal content and mesenteric lymph nodes were recovered. To determine bacterial loads in the mesenteric lymph node, the whole node was homogenized in PBS (0.5% tergitol, 0.5% bovine serum albumin). Minimal detectable values were 10 CFU/g in fecal and cecal content and 10 CFU/organ in the mesenteric lymph node. Parts of the cecal tissue were embedded in OCT (Sakura), and cryosections were prepared and stained with hematoxiline/eosine for pathoscoring. Evaluating submucosal edema, PMN infiltration, presence of goblet cells, and epithelial damage yielded a total score of 0–13 points as described (Hapfelmeier et al., 2008).

Precolonization Experiments

Bacterial strains for precolonization (*S.* Tm^{avir}, *S.* Tm$^{avir\ hyb}$) were grown for 12 hr at 37°C in LB supplemented with 0.3 M NaCl, diluted 1:20 into fresh medium, and subcultured for 4 hr. Animals starved for 4 hr were inoculated with 5×10^7 bacteria by gavage. Twenty-four hours later, fecal pellets were collected to check for successful colonization by plating ($\geq 10^7$ cfu/g feces), and animals were infected with a 1:1 mixture of *S.* Tm^{avir} and *S.* Tm$^{avir\ hyd3}$. Animals were sacrificed 24 hr later, and C.I.s were determined as described above.

In Vivo Screening-type Experiment

Library Generation

The transposon mutant library in *S.* Tm^{WT} was generated as previously described (Chan et al., 2005). Briefly, the suicide plasmid pJA1 (Badarinarayana et al., 2001) was mobilized from *E. coli* SM10 λ*pir* into SL1344 by conjugation for 6 hr in the presence of isopropyl-β-D-thiogalactopyranoside (IPTG) without antibiotic selection. During this time, the plasmid-encoded Tn10 transposase under control of an IPTG-inducible promoter is expressed. The mating reaction was harvested, and dilutions were plated on agar containing

Table 1. Microbiota Metagenomes Show Evidence for H_2-Producing Proteins

	FeFe Hydrogenase		NiFe Hydrogenase		Data Set	Sample Size
Hosts	Small Subunit PF02256	Large Subunit PF02906	Small Subunit PF14720	Large Subunit PF00374	Identifier	Total
Termite	+	+	–	+	Termite	165
Human	+	+	+	+	MetaHit	124
	+	+	+	+	AgeGeo	111
Mouse	–	+	–	–	Lean	1
	–	+	–	–	Obese	1
	+	+	+	+	LCM	1
Dog	+	+	–	+	K9C	6
	+	+	–	–	K9BP	6
Cow	+	+	–	+	Heifer	6
Chicken	–	+	–	–	A	1
	+	+	–	–	B	1

Metagenomes from six different species were analyzed for the presence of large and small subunit genes of FeFe- and NiFe-hydrogenases (Experimental Procedures; for further details, see Table S4). NiFe-hydrogenases comprise both H_2-consuming members and H_2-producing members. In contrast, the FeFe-hydrogenases generally produce (*not* consume) H_2 under anaerobic conditions and are therefore an indicator for hydrogen production within a microbial community (Schwartz and Friedrich, 2006). HmD-like enzymes were not considered, as they are only found in some methanogenic archaea. MG-Rast IDs, 44427013 (termite), 4440285 (chicken cecum A), 4440286 (chicken cecum B), 4444164 (canine K9c), 4444165 (canine K9bp), 4440463 (lean mouse), 4440464 (obese mouse), 4535626.3 and 4535627.3 (LCM mouse), 4448367.3 (cow), http://gutmeta.genomics.org.cn (MetaHIT human gut metagenome study), and 4461119-4461229 (human gut metagenome, "AgeGeo" study). See also Table S4.

200 μg/ml streptomycin and 30 μg/ml kanamycin to select for transposon-containing SL1344 bacteria. Single transposon insertion events per bacterial cells were checked by Southern blot with a probe directed against the transposon sequence (data not shown), and pools of 500 transposon mutants were stocked in peptone (5% glycerol) at −80°C.

Experimental Procedure

The screening-type experiment was adapted from the TraDIS (transposon differential insertion site sequencing) approach which was described previously (Chaudhuri et al., 2009, 2013). Six mice (two independent experiments of three animals each) were infected with a mix containing the pool of 500 transposon mutants and four wild-type isogenic tagged strains (WITS) (Grant et al., 2008) spiked in at a dilution of 1:500 (5×10^7 cfu total in 50 μl PBS). The spiked-in WITS strains contain a 40 nt barcode tag between the two pseudogenes *malX* and *malY* and allowed to check for random loss of subpopulations during the in vivo selection. An aliquot of the inoculum was grown up in LB broth (30 μg/ml kanamycin) and harvested as input pool. Animals were sacrificed at day 4 after infection. Cecal content was harvested, homogenized, and cultured overnight in LB (30 μg/ml kanamycin) to isolate transposon-containing output bacteria and in LB (12 μg/ml tetracycline) to isolate WITS-tagged strains for WITS analysis. Genomic DNA was prepared from input and output samples and fragmented, and RNA was amplified from the gDNA fragments using T7 RNA polymerase. Preparation of 5′ fragment cDNA libraries for 454 Titanium sequencing on a Roche/454 GS FLX sequencer (ca. 450 bp read length) was performed by vertis Biotechnologie AG (Freising, Germany). Briefly, RNA samples were poly(A)-tailed using poly(A) polymerase. An oligo(dT)-adaptor primer and M-MLV-H^- reverse transcriptase was used for first-strand cDNA synthesis. cDNA was amplified with PCR using primers directed to the flanking 5′ transposon and 3′ adaptor primer sequences and a proofreading enzyme. The double-stranded cDNA fragments then had a size of about 200–1,200 bp, were purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics), and were pooled for sequencing.

WITS Analysis

Temporal dynamics of WITS strains during screening experiments were monitored as described previously (Grant et al., 2008). In summary, WITS-tagged bacteria were harvested from enrichment cultures from fecal samples at day 1 after infection or cecum content samples at day 4 postinfection by centrifugation. Genomic bacterial DNA was extracted via the QIAGEN DNA mini kit, and the relative numbers of the four different WITS were determined by real-time PCR quantification using tag-specific primers.

Bioinformatic and Statistical Analysis of the 454 Sequencing Reads

The sequencing vendor provided reads split by barcode for the first sequencing run and pooled reads for the second sequencing run. The pooled sequences were split using a custom python script, using a perfect match criterion to the barcode sequences required. Transposon sequences were trimmed from the reads using Cutadapt version 1.1 (http://journal.embnet.org/index.php/embnetjournal/article/view/200), with a maximum error rate of 10%. The transposon sequence was detected (at least 92% of the reads) in each sample and removed. Untrimmed reads were discarded. Reads were mapped to the SL1344 genome (GenBank entry FQ312003.1) with Bowtie2 (http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html) version 2.0.0-beta6 using the –local parameter combination for local, gapped alignment, and sorted and converted to bam format using Samtools (http://bioinformatics.oxfordjournals.org/content/25/16/2078.short). Mapping start sites were counted using pysam (http://code.google.com/p/pysam/). Mapped reads starting within several nucleotides of each other were considered to belong to the same transposon insertion site. For each run of contiguous read start sites, the site with the highest coverage was chosen, and the total read count was calculated as the sum of the countiguous reads. Differential representation of the start sites between the input and output samples was estimated using edgeR (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/), using the generalized linear model framework (http://www.ncbi.nlm.nih.gov/pubmed/22287627) with tagwise dispersions. Counts per million were summed across samples, and start sites with a summed count equal to or less than 25 were excluded. The 30 most significantly attenuated start sites located within operon reading frames were selected for further analysis. Start sites overlapping a gene were annotated.

Lipocalin-2 ELISA

Lipocalin-2 levels were detected in homogenized fecal samples by ELISA using the DuoSet ELISA kit (R&D Systems).

Measurements of Cecal H_2 Concentration Using Clarke-type Microelectrodes

Hydrogen concentrations within the cecal lumen of mice with different microbiotas (CON, LCM, and GF) were measured using microsensors (Unisense, Aarhus, Denmark). The hydrogen microsensor (H-50) with a tip diameter of 50 μm was calibrated in water flushed with a gas mix containing 7% hydrogen at 37°C. This corresponds to a hydrogen concentration of 48.5 μM (Wiesenburg and Guinasso, 1979). Mice were sacrificed; ceca including ileum and

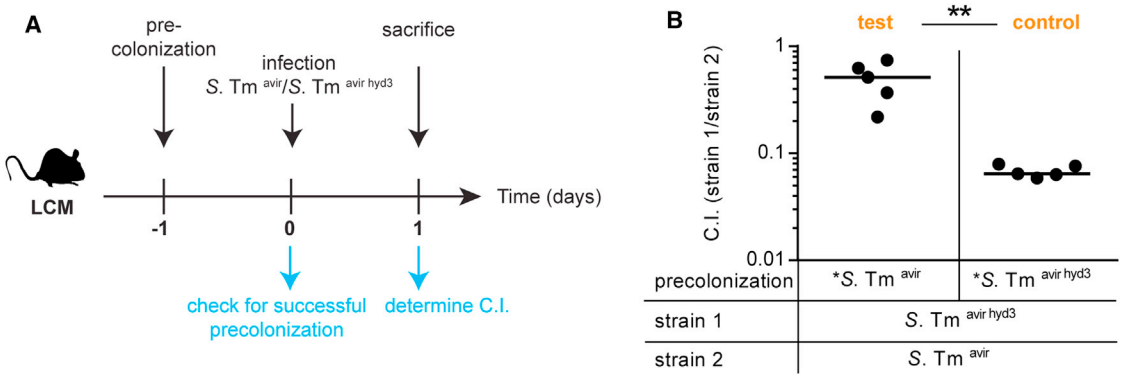


Figure 6. Introducing a Hydrogen Consumer Interferes with *hyb*-Dependent Gut Ecosystem Invasion by *S.* Tm

(A) Experimental strategy.

(B) LCM mice were precolonized with the hydrogen consumer *S.* Tm^{avir} (test) or a mutant incapable to consume hydrogen *S.* $Tm^{avir\ hyd3}$ (control; 5×10^7 cfu by gavage 1 day before infection). Plating verified the precolonization efficiency. Mice were infected with a 1:1 mixture of *S.* Tm^{avir} and *S.* $Tm^{avir\ hyd3}$ (5×10^7 cfu by gavage; five mice per group). C.I.s were determined at day 1 p.i. by differential plating of feces. $^{**}p < 0.01$, Mann-Whitney U test. Asterisk denotes that strains with distinct resistance markers were used for precolonization and for competitive infections.

See also Figure S5.

large intestine were fixed onto a bottom layer of 2% agarose in a petri dish and covered with top agar (45°C, 2% agarose) to fix the intestine as described (Schauer et al., 2012). A 26 G needle was used to pierce holes into the tissue to facilitate the microsensor tip to penetrate into the cecal lumen. After solidification of the top agar, the petri dish was transferred into a 37°C water bath, and microsensor profiles were taken at the prepierced positions. We measured three different spots per cecum: one at the cecal tip, one in the mid-cecum, and one at the opening toward small and large intestine. Please note that the values obtained by this method might be a bit higher than the steady-state levels in the gut of a living animal, as H_2 production is in equilibrium not only with microbial H_2 consumption but also with tissue diffusion, blood-mediated transport, and loss in breath and flatus (Bond and Levitt, 1972; Cummings and Macfarlane, 1991; Levitt et al., 1987).

To exclude artifacts attributable to H_2S, we performed measurements of hydrogen sulfide in parallel in the same mice at the same spots. The H_2S microsensor (H_2S-50) with a tip size of 50 μm was calibrated using an anaerobically prepared stock solution of S^{2-} (~0.01M). The final concentration of the stock solution was determined photometrically as previously described (Siegel, 1965). The H_2S microsensor detects the partial pressure of H_2S gas, a component of the total sulfide equilibrium system. At pH below 4, the equilibrium is shifted in favor of the gas, and all sulfides exist as gaseous H_2S. Therefore, the stock solution was diluted with degased technical buffer pH 1. Calibration values were taken at 37°C by removing the rubber stopper from the diluted calibration solutions (10 μM, 50 μM, and 200 μM), and the microsensor tip was immersed into the solution. We measured a median of 170 μM for CON mice, 63 μM for LCM mice, and 0 μM GF mice. Using these values, we corrected the signals measured with the H_2 microsensor for H_2S interference based on a crosssensititvity of 10% reported by the supplier (Unisense).

Metagenomic Analysis

DNA extraction of microbiota from murine feces of an LCM mouse of our colony was performed in the same way as for 16S rRNA gene sequencing (Supplemental Experimental Procedures). DNA library construction and high-throughput sequencing of the LCM microbiota metagenome were performed by BGI (Shenzhen, China) using Illumina's Hiseq technology (91PE) as previously decribed (Qin et al., 2010). The contigs were assembled using velvet with a k-mer length of 29, and host genomic sequences were filtered out using Bowtie2 and deposited as MG-Rast accession numbers 4535626.3 and 4535627.3.

Other sequences were retrieved from the public databases (Table 1). Nucleotide contig sets of the metagenomic data sets were procured from MG-RAST. These contig sets were prefiltered to remove the host genomic sequences. A six-frame translation was carried out on each of the individual data sets to identify any open reading frames coding for peptides longer than 30 amino acids. Next, a set of four pfam models—PF00374, PF02256, PF14720, and PF02906—was used for identifying homologs of hydrogenase subunits in our data sets. The initial screening was performed using Hmmscan with an value restriction of 0.0001, and these hits were reverse-screened against the entire Pfam HMM database.

Statistical Analysis

The one-sided Wilcoxon matched-pairs signed rank test and the exact Mann-Whitney U test were performed using the software Graphpad Prism Version 6.0 for Windows (GraphPad Software, http://www.graphpad.com). p values of less than 0.05 (two-tailed) were considered as statistically significant. To compare C.I.s to C.I. of inoculi, ratios of strain 1 and strain 2 were compared to the ratio of both strains in the inoculum using an exact Mann-Whitney U test.

Ethical Statement

All animal experiments were reviewed and approved by the Kantonales Veterinäramt, Zürich (license 223/2010 + Ergänzung 9) and are subject to the Swiss animal protection law (TschG).

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, four tables, and Supplemental Experimental Procedures and can be found with this article at http://dx.doi.org/10.1016/j.chom.2013.11.002.

ACKNOWLEDGMENTS

We are grateful to the members of the Hardt lab; to Tobias Erb, Andrew Macpherson, Julia Vorholt, and Hauke Hennecke for helpful scientific discussions; to Hans-Joachim Ruscheweyh (Center for Bioinformatics, Tübingen University) for support in 16S sequencing data analysis; to Thomas C. Weber and the RCHCI team (especially Corina Fusaro-Graf and Marion Hermerschmidt) for expert assistance with animal work; and to Manja Barthel and Maria Rita Lecca (FGCZ) for excellent technical support. This work was supported in part by the Swiss National Science Foundation (310030-132997/1 and the Sinergia project CRSII3_136286 to W.-D.H.).

Received: October 8, 2013
Revised: November 1, 2013
Accepted: November 11, 2013
Published: December 11, 2013

REFERENCES

Ackermann, M., Stecher, B., Freed, N.E., Songhet, P., Hardt, W.D., and Doebeli, M. (2008). Self-destructive cooperation mediated by phenotypic noise. Nature *454*, 987–990.

Allen-Vercoe, E., and Woodward, M.J. (1999). Colonisation of the chicken caecum by afimbriate and aflagellate derivatives of Salmonella enterica serotype Enteritidis. Vet. Microbiol. *69*, 265–275.

Badarinarayana, V., Estep, P.W., 3rd, Shendure, J., Edwards, J., Tavazoie, S., Lam, F., and Church, G.M. (2001). Selection analyses of insertional mutants using subgenic-resolution arrays. Nat. Biotechnol. *19*, 1060–1065.

Barthel, M., Hapfelmeier, S., Quintanilla-Martínez, L., Kremer, M., Rohde, M., Hogardt, M., Pfeffer, K., Rüssmann, H., and Hardt, W.D. (2003). Pretreatment of mice with streptomycin provides a Salmonella enterica serovar Typhimurium colitis model that allows analysis of both pathogen and host. Infect. Immun. *71*, 2839–2858.

Bond, J.H., Jr., and Levitt, M.D. (1972). Use of pulmonary hydrogen (H 2) measurements to quantitate carbohydrate absorption. Study of partially gastrectomized patients. J. Clin. Invest. *51*, 1219–1225.

Carbonero, F., Benefiel, A.C., and Gaskins, H.R. (2012). Contributions of the microbial hydrogen economy to colonic homeostasis. Nat. Rev. Gastroenterol. Hepatol. *9*, 504–518.

Chan, K., Kim, C.C., and Falkow, S. (2005). Microarray-based detection of Salmonella enterica serovar Typhimurium transposon mutants that cannot survive in macrophages and mice. Infect. Immun. *73*, 5438–5449.

Chaudhuri, R.R., Peters, S.E., Pleasance, S.J., Northen, H., Willers, C., Paterson, G.K., Cone, D.B., Allen, A.G., Owen, P.J., Shalom, G., et al. (2009). Comprehensive identification of Salmonella enterica serovar typhimurium genes required for infection of BALB/c mice. PLoS Pathog. *5*, e1000529.

Chaudhuri, R.R., Morgan, E., Peters, S.E., Pleasance, S.J., Hudson, D.L., Davies, H.M., Wang, J., van Diemen, P.M., Buckley, A.M., Bowen, A.J., et al. (2013). Comprehensive assignment of roles for Salmonella typhimurium genes in intestinal colonization of food-producing animals. PLoS Genet. *9*, e1003456.

Cotillard, A., Kennedy, S.P., Kong, L.C., Prifti, E., Pons, N., Le Chatelier, E., Almeida, M., Quinquis, B., Levenez, F., Galleron, N., et al.; ANR MicroObes Consortium (2013). Dietary intervention impact on gut microbial gene richness. Nature *500*, 585–588.

Craven, S.E. (1994). Altered colonizing ability for the ceca of broiler chicks by lipopolysaccharide-deficient mutants of Salmonella typhimurium. Avian Dis. *38*, 401–408.

Cummings, J.H., and Macfarlane, G.T. (1991). The control and consequences of bacterial fermentation in the human colon. J. Appl. Bacteriol. *70*, 443–459.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. USA *97*, 6640–6645.

Dewhirst, F.E., Chien, C.C., Paster, B.J., Ericson, R.L., Orcutt, R.P., Schauer, D.B., and Fox, J.G. (1999). Phylogeny of the defined murine microbiota: altered Schaedler flora. Appl. Environ. Microbiol. *65*, 3287–3292.

Endt, K., Stecher, B., Chaffron, S., Slack, E., Tchitchek, N., Benecke, A., Van Maele, L., Sirard, J.C., Mueller, A.J., Heikenwalder, M., et al. (2010). The microbiota mediates pathogen clearance from the gut lumen after non-typhoidal Salmonella diarrhea. PLoS Pathog. *6*, e1001097.

Fischbach, M.A., and Sonnenburg, J.L. (2011). Eating for two: how metabolism establishes interspecies interactions in the gut. Cell Host Microbe *10*, 336–347.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat. Rev. Microbiol. *6*, 121–131.

Food and Agriculture Organization of the United Nations (2002). W.H.O. (Risk Assessments of Salmonella in Eggs and Broiler Chickens).

Grant, A.J., Restif, O., McKinley, T.J., Sheppard, M., Maskell, D.J., and Mastroeni, P. (2008). Modelling within-host spatiotemporal dynamics of invasive bacterial disease. PLoS Biol. *6*, e74.

Hapfelmeier, S., Stecher, B., Barthel, M., Kremer, M., Müller, A.J., Heikenwalder, M., Stallmach, T., Hensel, M., Pfeffer, K., Akira, S., and Hardt, W.D. (2005). The Salmonella pathogenicity island (SPI)-2 and SPI-1 type III secretion systems allow Salmonella serovar typhimurium to trigger colitis via MyD88-dependent and MyD88-independent mechanisms. J. Immunol. *174*, 1675–1685.

Hapfelmeier, S., Müller, A.J., Stecher, B., Kaiser, P., Barthel, M., Endt, K., Eberhard, M., Robbiani, R., Jacobi, C.A., Heikenwalder, M., et al. (2008). Microbe sampling by mucosal dendritic cells is a discrete, MyD88-independent step in DeltainvG S. Typhimurium colitis. J. Exp. Med. *205*, 437–450.

Hoiseth, S.K., and Stocker, B.A. (1981). Aromatic-dependent Salmonella typhimurium are non-virulent and effective as live vaccines. Nature *291*, 238–239.

Ilg, K., Endt, K., Misselwitz, B., Stecher, B., Aebi, M., and Hardt, W.D. (2009). O-antigen-negative Salmonella enterica serovar Typhimurium is attenuated in intestinal colonization but elicits colitis in streptomycin-treated mice. Infect. Immun. *77*, 2568–2575.

Kaiser, P., Diard, M., Stecher, B., and Hardt, W.D. (2012). The streptomycin mouse model for Salmonella diarrhea: functional analysis of the microbiota, the pathogen's virulence factors, and the host's mucosal immune response. Immunol. Rev. *245*, 56–83.

Kerr, A.K., Farrar, A.M., Waddell, L.A., Wilkins, W., Wilhelm, B.J., Bucher, O., Wills, R.W., Bailey, R.H., Varga, C., McEwen, S.A., and Rajić, A. (2013). A systematic review-meta-analysis and meta-regression on the effect of selected competitive exclusion products on Salmonella spp. prevalence and concentration in broiler chickens. Prev. Vet. Med. *111*, 112–125.

Lamichhane-Khadka, R., Kwiatkowski, A., and Maier, R.J. (2010). The Hyb hydrogenase permits hydrogen-dependent respiratory growth of Salmonella enterica serovar Typhimurium. MBio. *1*, e00284-10.

Lamichhane-Khadka, R., Frye, J.G., Porwollik, S., McClelland, M., and Maier, R.J. (2011). Hydrogen-stimulated carbon acquisition and conservation in Salmonella enterica serovar Typhimurium. J. Bacteriol. *193*, 5824–5832.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S., et al.; MetaHIT consortium (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546.

Levitt, M.D., Hirsh, P., Fetzer, C.A., Sheahan, M., and Levine, A.S. (1987). H2 excretion after ingestion of complex carbohydrates. Gastroenterology *92*, 383–389.

Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R., and Gordon, J.I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. *6*, 776–788.

Maier, R.J. (2003). Availability and use of molecular hydrogen as an energy substrate for Helicobacter species. Microbes Infect. *5*, 1159–1163.

Maier, R.J. (2005). Use of molecular hydrogen as an energy substrate by human pathogenic bacteria. Biochem. Soc. Trans. *33*, 83–85.

Maier, R.J., Fu, C., Gilbert, J., Moshiri, F., Olson, J., and Plaut, A.G. (1996). Hydrogen uptake hydrogenase in Helicobacter pylori. FEMS Microbiol. Lett. *141*, 71–76.

Maier, R.J., Olczak, A., Maier, S., Soni, S., and Gunn, J. (2004). Respiratory hydrogen use by Salmonella enterica serovar Typhimurium is essential for virulence. Infect. Immun. *72*, 6294–6299.

McNorton, M.M., and Maier, R.J. (2012). Roles of H2 uptake hydrogenases in Shigella flexneri acid tolerance. Microbiology *158*, 2204–2212.

Ng, K.M., Ferreyra, J.A., Higginbottom, S.K., Lynch, J.B., Kashyap, P.C., Gopinath, S., Naidu, N., Choudhury, B., Weimer, B.C., Monack, D.M., and Sonnenburg, J.L. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. Nature *502*, 96–99.

Olson, J.W., and Maier, R.J. (2002). Molecular hydrogen as an energy source for Helicobacter pylori. Science *298*, 1788–1790.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Schauer, C., Thompson, C.L., and Brune, A. (2012). The bacterial community in the gut of the Cockroach Shelfordella lateralis reflects the close evolutionary relatedness of cockroaches and termites. Appl. Environ. Microbiol. *78*, 2758–2767.

Schwartz, E., and Friedrich, B. (2006). The H2-metabolizing prokaryoes. In The Prokaryotes: A Handbook on the Biology of Bacteria, Vol. 2, Ecophysiology and Biochemistry, M. Dworkin, ed. (New York: Springer).

Siegel, L.M. (1965). A Direct Microdetermination for Sulfide. Anal. Biochem. *11*, 126–132.

Sommer, F., and Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. Nat. Rev. Microbiol. *11*, 227–238.

Stecher, B., and Hardt, W.D. (2011). Mechanisms controlling pathogen colonization of the gut. Curr. Opin. Microbiol. *14*, 82–91.

Stecher, B., Hapfelmeier, S., Müller, C., Kremer, M., Stallmach, T., and Hardt, W.D. (2004). Flagella and chemotaxis are required for efficient induction of Salmonella enterica serovar Typhimurium colitis in streptomycin-pretreated mice. Infect. Immun. *72*, 4138–4150.

Stecher, B., Robbiani, R., Walker, A.W., Westendorf, A.M., Barthel, M., Kremer, M., Chaffron, S., Macpherson, A.J., Buer, J., Parkhill, J., et al. (2007). Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. PLoS Biol. *5*, 2177–2189.

Stecher, B., Barthel, M., Schlumberger, M.C., Haberli, L., Rabsch, W., Kremer, M., and Hardt, W.D. (2008). Motility allows S. Typhimurium to benefit from the mucosal defence. Cell. Microbiol. *10*, 1166–1180.

Stecher, B., Chaffron, S., Käppeli, R., Hapfelmeier, S., Freedrich, S., Weber, T.C., Kirundi, J., Suar, M., McCoy, K.D., von Mering, C., et al. (2010). Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. PLoS Pathog. *6*, e1000711.

van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat. Rev. Microbiol. *11*, 435–442.

Wiesenburg, D.A., and Guinasso, N.L. (1979). Equilibrium solubilities of methane, carbon monoxide and hydrogen in water and sea water. J. Chem. Eng. Data *24*, 356–360.

Willing, B.P., Russell, S.L., and Finlay, B.B. (2011). Shifting the balance: antibiotic effects on host-microbiota mutualism. Nat. Rev. Microbiol. *9*, 233–243.

Winter, S.E., Lopez, C.A., and Bäumler, A.J. (2013). The dynamics of gut-associated microbial communities during inflammation. EMBO Rep. *14*, 319–327.

Zbell, A.L., and Maier, R.J. (2009). Role of the Hya hydrogenase in recycling of anaerobically produced H2 in Salmonella enterica serovar Typhimurium. Appl. Environ. Microbiol. *75*, 1456–1459.

Zbell, A.L., Benoit, S.L., and Maier, R.J. (2007). Differential expression of NiFe uptake-type hydrogenase genes in Salmonella enterica serovar Typhimurium. Microbiology *153*, 3508–3516.

Zbell, A.L., Maier, S.E., and Maier, R.J. (2008). Salmonella enterica serovar Typhimurium NiFe uptake-type hydrogenases are differentially expressed in vivo. Infect. Immun. *76*, 4445–4454.

8.2 List of abbreviations

16S	16S small subunit ribosomal RNA gene
al	**a**verage **l**inkage hierarchical clustering
AMI	**A**djusted **M**utual **I**nformation
ARI	**A**djusted **R**and **I**ndex
BER	**b**road **e**cological **r**ange 'local' SSU dataset
Chao1	Chao1 community richness estimator
cl	**c**omplete **l**inkage hierarchical clustering
ECS	**E**cological **C**onsistency **S**core
HCA	**h**ierarchical **c**lustering **a**lgorithm
HMP	**h**uman **m**icrobiome **p**roject
HSM	**h**uman **s**kin **m**icrobiome SSU dataset [51]
J_{abd}	**J**accard index, **ab**un**d**ance-corrected after Chao, 2004 [155]
MI	**m**utual **i**nformation
MH	**M**orisita-**H**orn overlap index
MSA	**m**ultiple **s**equence **a**lignment
NMI	**N**ormalized **M**utual **I**nformation
nt	**n**ucleo**t**ides (as a measure of sequence length)
OTU	**O**perational **T**axonomic **U**nit
PCR	**p**olymerase **c**hain **r**eaction
PSA	**p**airwise **s**equence **a**lignment
RDP	**R**ibosomal **D**atabase **P**roject
SDC	**S**ørensen-**D**ice-**C**zechanowski index of community similarity
Shannon	Shannon entropy, as index of community richness
sl	**s**ingle **l**inkage hierarchical clustering
SLP	**s**ingle **l**inkage **p**re-clustering
SSU	**s**mall **su**bunit ribosomal RNA gene
V23, V35 and V6	hyper-**v**ariable subregions 2-3, 3-5 and 6 of the SSU rRNA gene
VI	**V**ariation of **I**nformation

8.2. References

1. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci USA 74: 4537–4541.

2. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA 74: 5088–5090. doi:10.1073/pnas.74.11.5088.

3. Shine J, Dalgarno L (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci USA 71: 1342–1346.

4. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, et al. (2000) Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. Cell 102: 615–623. doi:10.1016/S0092-8674(00)00084-2.

5. de Peer Van Y, Chapelle S, De Wachter R (1996) A Quantitative Map of Nucleotide Substitution Rates in Bacterial rRNA. Nucleic Acids Research 24: 3381–3391. doi:10.1093/nar/24.17.3381.

6. Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271.

7. Fox GE, Wisotzkey JD, Jurtshuk P (1992) How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. International Journal of Systematic Bacteriology 42: 166–170. doi: 10.1099/00207713-42-1-166.

8. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124–2129.

9. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, et al. (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3: 733–739. doi:doi:10.1038/nrmicro1236.

10. Stackebrandt E, Goebel BM (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol 44: 846–849. doi: 10.1099/00207713-44-4-846.

11. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, et al. (2008) The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31: 241–250. doi:10.1016/j.syapm.2008.07.001.

12. de Wit R, Bouvier T (2006) "Everything is everywhere, but, the environment selects;" what did Baas Becking and Beijerinck really say? Environ Microbiol 8: 755–758. doi:10.1111/j.1462-2920.2006.01017.x.

13. Dworkin M (2012) Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. FEMS Microbiol Rev 36: 364–379. doi:10.1111/j.1574-6976.2011.00299.x.

14. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci USA 82: 6955–6959.

15. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321–346. doi:10.1146/annurev.mi.39.100185.001541.

16. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: reviews0003.1. doi: 10.1186/gb-2002-3-2-reviews0003.

17. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638. doi:10.1126/science.1110591.

18. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5: R245–R249. doi:10.1016/S1074-5521(98)90108-9.

19. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669–685. doi:10.1128/MMBR.68.4.669-685.2004.

20. Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38: 525–552. doi:10.1146/annurev.genet.38.072902.091216.

21. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, et al. (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Bacteriol 37: 463–464. doi: 10.1099/00207713-37-4-463.

22. Stackebrandt E, Frederiksen W, Garrity GM, Grimont P, Kampfer P, et al. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52: 1043–1047. doi: 10.1099/ijs.0.02360-0.

23. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, et al. (2013) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research 42: D643–D648. doi:10.1093/nar/gkt1209.

24. Hugenholtz P, Goebel BM, Pace NR (1998) Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J Bacteriol 180: 4765. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC107498/.

25. Stackebrandt E (2009) Phylogeny Based on 16S rRNA/DNA. Encyclopedia of Life Sciences. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1038/npg.els.0000462.

26. Olsen GJ, Overbeek R, Larsen N, Marsh TL, McCaughey MJ, et al. (1992) The Ribosomal Database Project. Nucleic Acids Research 20: 2199–2200. doi:10.1093/nar/20.suppl.2199.

27. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, et al. (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research 42: D633–D642. doi:10.1093/nar/gkt1244.

28. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072. doi:10.1128/AEM.03006-05.

29. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35: 7188–7196. doi:10.1093/nar/gkm864.

30. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380. doi:10.1038/nature03959.

31. Bennett S (2004) Solexa Ltd. Pharmacogenomics 5: 433–438. doi:10.1517/14622416.5.4.433.

32. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463.

33. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotech 26: 1135–1145. doi:doi:10.1038/nbt1486.

34. Consortium THMP (2012) A framework for human microbiome research. Nature 486: 215–221. doi:doi:10.1038/nature11209.

35. Gilbert J, Meyer F, Antonopoulos DA, Balaji P, Brown CT, et al. (2010) Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand Genomic Sci 3: 243–248. doi:10.4056/sigs.1433550.

36. V Wintzingerode F, Göbel UB, Stackebrandt E (2006) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21: 213–229. doi:10.1111/j.1574-6976.1997.tb00351.x.

37. Patin NV, Kunin V, Lidström U, Ashby MN (2012) Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. Microb Ecol 65: 709–719. doi:10.1007/s00248-012-0145-4.

38. Meyerhans A, Vartanian J-P, Wain-Hobson S (1990) DNA recombination during PCR. Nucleic Acids Research 18: 1687.

39. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. Appl Environ Microbiol 71: 7724–7736. doi:10.1128/AEM.71.12.7724-7736.2005.

40. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21: 494–504. doi:10.1101/gr.112730.110.

41. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics (Oxford, England) 27: 2194–2200. doi:10.1093/bioinformatics/btr381.

42. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143. doi:10.1186/gb-2007-8-7-r143.

43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12: 2947–2948. Available: http://bioinformatics.oxfordjournals.org/content/23/21/2947.long.

44. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotech 30: 434–439. doi:10.1038/nbt.2198.

45. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. PLOS ONE 7: e30087. doi:10.1371/journal.pone.0030087.

46. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Publishing Group 6: 639–641. doi:10.1038/nmeth.1361.

47. Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ (2011) Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics 12: 38. doi:10.1186/1471-2105-12-38.

48. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, et al. (2012) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Publishing Group 10: 57–59. doi:10.1038/nmeth.2276.

49. Li W, Fu L, Niu B, Wu S, Wooley J (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. Briefings in bioinformatics 13: 656–668. doi:10.1093/bib/bbs035.

50. Schloss PD (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLOS Computational biology 6: e1000844. doi:10.1371/journal.pcbi.1000844.

51. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, et al. (2009) Topographical and Temporal Diversity of the Human Skin Microbiome. Science 324: 1190–1192. doi:10.1126/science.1171700.

52. Wetterstrand KA (n.d.) DNA Sequencing Costs. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available: http://www.genome.gov/sequencingcosts/. Accessed 28 March 2014.

53. Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol 12: 1889–1898. doi:10.1111/j.1462-2920.2010.02193.x.

54. Sneath PHA, Sokal RR (1973) Numerical Taxonomy. W H Freeman & Company. 1 pp.

55. Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLOS Computational biology 6: e1000667. doi:10.1371/journal.pcbi.1000667.t002.

56. Schloss PD, Handelsman J (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. Appl Environ Microbiol 71: 1501–1506. doi:10.1128/AEM.71.3.1501-1506.2005.

57. Schloss PD, Westcott SL (2011) Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. Appl Environ Microbiol 77: 3219–3226. doi:10.1128/AEM.02810-10.

58. Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, et al. (2011) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in bioinformatics 13: 107–121. doi:10.1093/bib/bbr009.

59. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol 73: 5261–5267. doi:10.1128/AEM.00062-07.

60. Schloss PD, Westcott SL, Rabyn T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75: 7537. doi:10.1128/AEM.01541-09.

61. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7: 335–336. doi:10.1038/nmeth.f.303.

62. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Res 8: 175–185. doi:10.1101/gr.8.3.175.

63. Quinlan AR, Stewart DA, Strömberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nature Methods 5: 179–181. doi:10.1038/nmeth.1172.

64. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nature Methods 7: 668–669. doi:10.1038/nmeth0910-668b.

65. Schloss PD, Gevers D, Westcott SL (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. PLOS ONE 6: e27310.

66. Bonder MJ, Abeln S, Zaura E, Brandt BW (2012) Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics (Oxford, England) 28: 2891–2897. doi:10.1093/bioinformatics/bts552.

67. Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics (Oxford, England) 20: 2317–2319. doi:10.1093/bioinformatics/bth226.

68. Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005) Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics (Oxford, England) 21: 333–337. doi:10.1093/bioinformatics/bti008.

69. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Publishing Group 10: 996–998. doi:10.1038/nmeth.2604.

70. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48: 443–453. doi:10.1016/0022-2836(70)90057-4.

71. Sun Y, Cai Y, Liu L, Yu F, Farrell ML, et al. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Research 37: e76–e76. doi:10.1093/nar/gkp285.

72. Cai Y, Sun Y (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Research 39: e95. doi:10.1093/nar/gkr349.

73. Wang X, Cai Y, Sun Y, Knight R, Mai V (2011) Secondary structure information does not improve OTU assignment for partial 16s rRNA sequences. ISME J 6: 1277–1280. doi:10.1038/ismej.2011.187.

74. Wang L, Jiang T (1994) On the Complexity of Multiple Sequence Alignment. J Comput Biol 1: 337–348. doi: 10.1089/cmb.1994.1.337.

75. Just W (2001) Computational Complexity of Multiple Sequence Alignment with SP-Score. J Comput Biol 8: 615–623. doi:10.1089/10665270175330751l.

76. Elias I (2006) Settling the Intractability of Multiple Alignment. J Comput Biol 13: 1323–1339. doi:10.1089/cmb.2006.13.1323.

77. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797. doi:10.1093/nar/gkh340.

78. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113. doi:10.1186/1471-2105-5-113.

79. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30: 772–780. doi:10.1093/molbev/mst010.

80. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics (Oxford, England) 25: 1335–1337. doi:10.1093/bioinformatics/btp157.

81. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics (Oxford, England) 29: 2933–2935. doi:10.1093/bioinformatics/btt509.

82. White JR, Navlakha S, Nagarajan N, Ghodsi M-R, Kingsford C, et al. (2010) Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. BMC Bioinformatics 11: 152. doi: 10.1186/1471-2105-11-152.

83. Barriuso J, Valverde JR, Mellado RP (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. BMC Bioinformatics 12: 473. doi:10.1186/1471-2105-12-473.

84. Schloss PD (2012) Secondary structure improves OTU assignments of 16S rRNA gene sequences. ISME J 7: 457–460. doi:10.1038/ismej.2012.102.

85. Matias Rodrigues JF, Mering von C (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. Bioinformatics (Oxford, England) 30: 287–288. doi:10.1093/bioinformatics/btt657.

86. Wei D, Jiang Q, Wei Y, Wang S (2012) A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics {13}. doi:{10.1186/1471-2105-13-174}.

87. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics (Oxford, England) 17: 282–283. doi:10.1093/bioinformatics/17.3.282.

88. Li W, Jaroszewski L, Godzik A (2002) Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Engineering Design and Selection 15: 643–649. doi:10.1093/protein/15.8.643.

89. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England) 22: 1658–1659. doi:10.1093/bioinformatics/btl158.

90. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England) 28: 3150–3152. doi:10.1093/bioinformatics/bts565.

91. Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. Protein Sci 1: 409–417. doi:10.1002/pro.5560010313.

92. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26: 2460–2461.

93. Lee J-H, Yi H, Jeon Y-S, Won S, Chun J (2012) TBC: A clustering algorithm based on prokaryotic taxonomy. J Microbiol 50: 181–185. doi:10.1007/s12275-012-1214-6.

94. Zheng Z, Kramer S, Schmidt B (2012) DySC: software for greedy clustering of 16S rRNA reads. Bioinformatics (Oxford, England) 28: 2182–2183. doi:10.1093/bioinformatics/bts355.

95. Chen W, Cheng Y, Zhang C, Zhang S, Zhao H (2013) MSClust: A Multi-Seeds based Clustering algorithm for microbiome profiling using 16S rRNA sequence. Journal of Microbiological Methods 94: 347–355. doi:10.1016/j.mimet.2013.07.004.

96. Rasheed Z, Rangwala H, Barbará D (2013) 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. BMC systems biology 7: S11. doi:10.1186/gb-2007-8-7-r143.

97. Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics (Oxford, England) 27: 611–618. doi:10.1093/bioinformatics/btq725.

98. Cheng L, Walker AW, Corander J (2012) Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Research 40: 5240–5249. doi:10.1093/nar/gks227.

99. Hwang K, Oh J, Kim T-K, Kim BK, Yu DS, et al. (2013) CLUSTOM: A Novel Method for Clustering 16S rRNA Next Generation Sequences by Overlap Minimization. PLOS ONE 8: e62623. doi:10.1371/journal.pone.0062623.

100. Wang X, Yao J, Sun Y, Mai V (2013) M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. BMC Bioinformatics 14: 43. doi:10.1093/nar/gks227.

101. Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, et al. (2011) PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data. PLOS Computational biology 7: e1001061. Available: http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1001061.

102. Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics (Oxford, England) 29: 2869–2876. doi:10.1093/bioinformatics/btt499.

103. Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, et al. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci USA 105: 2504–2509. doi:10.1073/pnas.0712205105.

104. Koeppel AF, Wu M (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Research 41: 5175–5188. doi:10.1093/nar/gkt241.

105. Cohan FM, Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. Current Biology 17: R373–R386. doi:10.1016/j.cub.2007.03.032.

106. Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ (2013) Distribution-Based Clustering: Using Ecology to Refine the Operational Taxonomic Unit. Appl Environ Microbiol 79: 6593–6603. doi:10.1128/AEM.00342-13.

107. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol 12: 118–123. doi:10.1111/j.1462-2920.2009.02051.x.

108. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 103: 12115–12120. doi:10.1073/pnas.0605127103.

109. Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H (2013) A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. PLOS ONE 8: e70837. doi:10.1371/journal.pone.0070837.s005.

110. Kim M, Morrison M, Yu Z (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. Journal of Microbiological Methods 84: 81–87. doi:10.1016/j.mimet.2010.10.020.

111. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484. doi:10.1038/nature07540.

112. Cohan FM (2002) What are bacterial species? Annu Rev Microbiol 56: 457–487. doi:10.1146/annurev.micro.56.012302.160634.

113. Doolittle WF, Zhaxybayeva O (2009) On the origin of prokaryotic species. Genome Res 19: 744–756. doi:10.1101/gr.086645.108.

114. Rosselló-Móra R (2011) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol 14: 318–334. doi:10.1111/j.1462-2920.2011.02599.x.

115. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430: 551–554. doi:10.1038/nature02649.

116. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. Science 307: 1311–1313. doi:10.1126/science.1106028.

117. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, et al. (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science 311: 1737–1740. doi:10.1126/science.1118052.

118. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, et al. (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proceedings of the National Academy of Sciences 108: 7200–7205. doi:10.1073/pnas.1015622108.

119. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, et al. (2012) Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336: 48–51. doi:10.1126/science.1218198.

120. Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison New York, NY: ACM Press. pp. 1073–1080. doi:10.1145/1553374.1553511.

121. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, et al. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320: 1081–1085. doi:10.1126/science.1157890.

122. Becraft ED, Cohan FM, Kühl M, Jensen SI, Ward DM (2011) Fine-scale distribution patterns of Synechococcus ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. Appl Environ Microbiol 77: 7689–7697. doi:10.1128/AEM.05927-11.

123. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Research 41: D36–D42. doi:10.1093/nar/gks1195.

124. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40: D130–D135. doi:10.1093/nar/gkr1079.

125. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596. doi:10.1093/nar/gks1219.

126. Chaffron S, Rehrauer H, Pernthaler J, Mering von C (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res 20: 947–959. doi:10.1101/gr.104521.109.

127. Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, et al. (2008) A diversity profile of the human skin microbiota. Genome Res 18: 1043–1050. doi:10.1101/gr.075549.107.

128. Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, et al. (2008) It's all relative: ranking the diversity of aquatic bacterial communities. Environ Microbiol 10: 2200–2210. doi:10.1111/j.1462-2920.2008.01626.x.

129. Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, et al. (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. Appl Environ Microbiol 74: 5422–5428. doi:10.1128/AEM.00410-08.

130. Brulc JM, Antonopoulos DA, Miller MEB, Wilson MK, Yannarell AC, et al. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. PNAS 106: 1948–1953. doi:10.1073/pnas.0806191105.

131. Alonso-Gutierrez J, Figueras A, Albaiges J, Jimenez N, Vinas M, et al. (2009) Bacterial Communities from Shoreline Environments (Costa da Morte, Northwestern Spain) Affected by the Prestige Oil Spill. Appl Environ Microbiol 75: 3407–3418. doi:10.1128/AEM.01776-08.

132. Cruz-Martínez K, Suttle KB, Brodie EL, Power ME, Andersen GL, et al. (2009) Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. ISME J 3: 738–744. doi:10.1038/ismej.2009.16.

133. Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, et al. (2009) Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science 326: 578–582. doi:10.1126/science.1175309.

134. Sunagawa S, Woodley CM, Medina M (2010) Threatened corals provide underexplored microbial habitats. PLOS ONE 5: e9554. doi:10.1371/journal.pone.0009554.

135. Durso LM, Harhay GP, Smith TPL, Bono JL, Desantis TZ, et al. (2010) Animal-to-animal variation in fecal microbial diversity among beef cattle. Appl Environ Microbiol 76: 4858–4862. doi:10.1128/AEM.00207-10.

136. Eloe EA, Shulse CN, Fadrosh DW, Williamson SJ, Allen EE, et al. (2010) Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. Environmental Microbiology Reports 3: 449–458. doi:10.1111/j.1758-2229.2010.00223.x.

137. Perkins SD, Angenent LT (2010) Potential pathogenic bacteria in metalworking fluids and aerosols from a machining facility. FEMS Microbiol Ecol 74: 643–654. doi:10.1111/j.1574-6941.2010.00976.x.

138. Martinson VG, Danforth BN, Minckley RL, Rueppell O, Tingek S, et al. (2011) A simple and distinctive microbiota associated with honey bees and bumble bees. Molecular Ecology 20: 619–628. doi:10.1111/j.1365-294X.2010.04959.x.

139. Perkins SD, Scalfone NB, Angenent LT (2011) Comparative 16S rRNA gene surveys of granular sludge from three upflow anaerobic bioreactors treating purified terephthalic acid (PTA) wastewater. Water Sci Technol 64: 1406–1412. doi:10.2166/wst.2011.552.

140. Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Research 40: D136–D143. doi:10.1093/nar/gkr1178.

141. Amante C, Atkins BW (2009) ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. National Oceanic and Atmospheric Administration. 25 pp.

142. Tuomisto H (2010) A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia 164: 853–860. doi:10.1007/s00442-010-1812-0.

143. Whittaker RH (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30: 279–338.

144. Whittaker RH (1972) Evolution and Measurement of Species Diversity. Taxon 21: 213–251.

145. Tuomisto H (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33: 2–22. doi:10.1111/j.1600-0587.2009.05880.x.

146. Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev 32: 557–578. doi:10.1111/j.1574-6976.2008.00111.x.

147. Lozupone C, Knight R (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol 71: 8228–8235. doi:10.1128/AEM.71.12.8228-8235.2005.

148. Chao A (1984) Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics 11: 265–270.

149. Chao A, Lee S-M (1992) Estimating the Number of Classes via Sample Coverage. Journal of the American Statistical Association 87: 210–217.

150. Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. Philos Trans R Soc Lond, B, Biol Sci 345: 101–118. doi:10.1098/rstb.1994.0091.

151. Shannon CE (1948) A Mathematical Theory of Communication. At&T Tech J 27: 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x.

152. Simpson EH (1949) Measurement of Diversity. Nature 163: 688–688. doi:10.1038/163688a0.

153. McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8: e61217. doi:10.1371/journal.pone.0061217.

154. Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37: 547–579.

155. Chao A, Chazdon RL, Colwell RK, Shen T-J (2004) A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett 8: 148–159. doi:10.1111/j.1461-0248.2004.00707.x.

156. Dice LR (1945) Measures of the Amount of Ecologic Association Between Species. Ecology 26: 297–302.

157. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 27: 325–349.

158. Horn HS (1966) Measurement of" overlap" in comparative ecological studies. American Naturalist: 419–424.

159. Meilă M (2007) Comparing clusterings—an information based distance. Journal of Multivariate Analysis 98: 873–895. doi:10.1016/j.jmva.2006.11.013.

160. Rand WM (1971) Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66: 846–850. doi:10.1080/01621459.1971.10482356.

161. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218. doi:10.1007/BF01908075.

162. Meilă M (2005) Comparing Clusterings: An Axiomatic View New York, NY, USA: ACM. pp. 577–584. doi: 10.1145/1102351.1102424.

163. Fred ALN, Jain AK (2003) Robust Data Clustering Vol. 3. pp. 128–136.

164. Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biol 7: 116. doi:10.1186/gb-2006-7-9-116.

165. Doolittle WF (2012) Population Genomics: How Bacterial Species Form and Why They Don't Exist. Current Biology 22: R451–R453. doi:10.1016/j.cub.2012.04.034.

166. Cohan FM (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos T R Soc B 361: 1985–1996. doi:10.1098/rstb.2006.1918.

167. Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6: 431–440. doi:10.1038/nrmicro1872.

168. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741–746. doi:10.1126/science.1159388.

169. Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19: 1–7. doi: 10.1016/j.tim.2010.10.003.

170. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, et al. (2010) The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol 8: 523–529. doi:10.1038/nrmicro2367.

171. Maughan H, Van der Auwera G (2011) Bacillus taxonomy in the genomic era finds phenotypes to be essential though often misleading. Infect Genet Evol 11: 789–797. doi:10.1016/j.meegid.2011.02.001.

172. Mering von C, Hugenholtz P, Raes J, Tringe SG, Doerks T, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315: 1126–1130. doi:10.1126/science.1133420.

173. Losos JB (2008) Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecol Lett 11: 995–1003. doi:10.1111/j.1461-0248.2008.01229.x.

174. Darwin C (1859) On the origin of species by means of natural selection. London: Murray.

8.3 Curriculum vitae

Name	Thomas Sebastian Benedikt Schmidt
Academic Title	Diplom-Ingenieur der Biotechnologie Master of Science, Ingénierie et Sciences de l'Environnement
Date of Birth	August 17th, 1987
Nationality	German
Current Position	PhD student, Institute of Molecular Life Sciences, UZH
Email	sebastian.schmidt@imls.uzh.ch

Education:

2005	Abitur, Annette von Droste-Hülshoff-Gymnasium Gelsenkirchen, Germany
2005-6	Studies in Biosciences Westfälische-Wilhelms-Universität Münster, Germany
2006-7	Studies in Biology Albert-Ludwigs-Universität Freiburg i. Br., Germany
2007	Vordiplom Biology
2007-2010	Studies in Biotechnology École Supérieure de Biotechnologie Strasbourg, Illkirch, France
2009-10	Studies in Environmental Sciences École et Observatoire des Sciences de la Terre, Strasbourg, France
2010	Diplôme d'Ingénieur en Biotechnologie École Supérieure de Biotechnologie Strasbourg, Illkirch, France Université de Strasbourg, France Albert-Ludwigs-Universität Freiburg i. Br., Germany Universität Karlsruhe, Germany Universität Basel, Switzerland Thesis: 'Metabolites as State Variables: Metabolic Flux Sensing in *E. coli*' Supervisor: Prof. Dr. Uwe Sauer, ETH Zürich
2010	Master of Science Ingénierie et Sciences de l'Environnement École et Observatoire des Sciences de la Terre, Strasbourg, France Université de Strasbourg, France
11/2010-present	PhD student, Institute of Molecular Life Sciences, UZH

8.4 Publication list

First-author manuscripts relevant to the present dissertation, reprinted in sections 7.1 & 7.2

Schmidt, T.S.B., Matias Rodrigues, J.F. & von Mering, C. (2014). **Limits to Robustness and Reproducibility in the Demarcation of Operational Taxonomic Units.** *under revision*

Schmidt, T.S.B., Matias Rodrigues, J.F. & von Mering, C. (2014). **Ecological Consistency of Operational Taxonomic Units at a Global Scale.** PLOS Computational Biology, 10(4), e1003594. doi:10.1371/journal.pcbi.1003594

Co-authored manuscript, reprinted in section 8.1

Maier, L., Vyas, R., Cordova, C. D., Lindsay, H., Schmidt, T. S. B., Brugiroux, S., et al. (2013). **Microbiota-Derived Hydrogen Fuels Salmonella Typhimurium Invasion of the Gut Ecosystem.** Cell Host & Microbe, 14(6), 641–651. doi:10.1016/j.chom.2013.11.002

Additional manuscripts, not reprinted or not relevant to the present dissertation

Pistohl, T., Schmidt, T. S. B., Ball, T., Schulze-Bonhage, A., Aertsen, A., & Mehring, C. (2013). **Grasp Detection from Human ECoG during Natural Reach-to-Grasp Movements.** PLOS One, 8(1), e54658. doi:10.1371/journal.pone.0054658

Maier, S., Schmidt, T.S.B., Zheng, L., Peer, T., Wagner, V. & Grube, M. (2014). **Specific Enrichment of Bacterial Communities in Lichens Forming Biological Soil Crusts.** Biodiversity & Conservation, doi:10.1007/s10531-014-0719-1

Becker, E.*, Schmidt, T.S.B.*, Stanzel, C., Atrott, K., Biedermann, L., Rehman, A., Jonas, D., von Mering, C., Rogler, G., Frey-Wagner, I. (2014). **Influence of Isotretinoin Treatment on the Murine Gastrointestinal Tract.** *in preparation*

A full and regularly updated list of published manuscripts can be found on my Google Scholar profile:

http://scholar.google.com/citations?user=E3pOqaEAAAAJ

8.5 Acknowledgements

Science is a social endeavor – and modern biology doubly so. An understanding of the world without is not achieved in the void: it takes the collective effort of many to wrest even the slightest secret from Gaia's reluctant bosom. Sort of. Less poetically put: scientists join forces to generate 'knowledge', and understanding may arise (if at all) from the constant friction generated by the pinball game of idea and counter-hypothesis. Even Isaac Newton, himself a leviathan of western science history, famously admitted to "standing on the shoulders of giants". How much more true is that of today's scientists.

They who aspire to accomplish an academic dissertation will experience the very same: they will find themselves on the shoulders of giants – or at least, faced with a giant body of research literature. And if they are lucky, they will be exposed to an environment of constant idea pinball. I have been lucky enough to work in such an environment for the past years, for which I sincerely thank the past and present members of the von Mering group at the University of Zürich. That, and for the innumerable cakes, coffee breaks, weird discussions and the exceptionally easy-going atmosphere that you do not find in many (research) groups.

There have been others who have allowed me to bounce off them, scientifically speaking: the members of my PhD committee, Kentaro Shimizu, Wolf-Dietrich Hardt, Bernhard Schmid and Jeroen Raes. They helped me more than they probably realize, in many different ways, and I am deeply thankful for their commitment.

I am thankful to João Rodrigues who really taught me almost everything I know – about ping-pong. He has also been the best postdoctoral colleague that a PhD student could possibly hope for: patient, highly knowledgeable, balanced, cooperative by nature and conviction, mindbogglingly fluent in at least five languages (including Polish, Swiss German, Cantonese and C/C++). João has been my colleague, mentor, travel companion and friend – for which I am deeply grateful.

I am thankful to my thesis supervisor Christian von Mering. When I first interviewed for a PhD position, an (undisclosed) group member told me plainly that "Christian is the best boss you could ever hope for." After three and a half years of first-hand experience I can quite simply confirm that in every possible way. I am genuinely glad to have been in his group, and I leave it at that: I am sincerely and deeply grateful.

Finally, if science does not happen in the void, neither does *life*. A doctorate, as so many other things in life, is fundamentally Pyrrhic if acquired at the cost of what is truly important. Although it is admittedly clichéd, I mean it: I am thankful to the people surrounding me – to my family.

And for more reasons than I can possibly enumerate, reasons that would lose their magic and meaning if committed to words on a dry piece of dissertation paper; for reasons that I cherish in my heart (where they belong) – I am grateful to Sina.